Evaluating the Applications of Spatial Audio in Telephony

by

Konrad Blum

Thesis presented in partial fulfilment of the requirements for the degree of

Master of Science in Engineering

at Stellenbosch University

Supervisor: Dr Gert-Jan van Rooyen

Co-supervisor: Dr Herman Arnold Engelbrecht

Department of Electrical & Electronic Engineering

March 2010

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the owner of the copyright thereof (unless to the extent explicitly otherwise stated) and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

March 2010

Abstract

Telephony has developed substantially over the years, but the fundamental auditory model of mixing all the audio from different sources together into a single monaural stream has not changed since the telephone was first invented. Monaural audio is very difficult to follow in a multiple-source situation such as a conference call.

Sound originating from a specific point in space will travel along a slightly different path to each ear. Although we are not consciously aware of it, our brain processes these spatial cues to help us to locate sounds in space. It is this spatial information that allows us to focus our attention and listen to a single speaker in an environment where many different sources may be active at the same time; a phenomenon known as the "cocktail party effect". It is possible to reproduce these spatial cues in a sound recording, using Head-Related Transfer Functions (HRTFs) to allow a listener to experience localised audio, even when sound is reproduced through a headset.

In this thesis, spatial audio is implemented in a telephony application as well as in a virtual world. Experiments were conducted which demonstrated that spatial audio increases the intelligibility of speech in a multiple-source environment and aids active speaker identification. Resource usage measurements show that these benefits are, however, not without a cost. In conclusion, spatial audio was shown to be an improvement over the monaural audio model traditionally implemented in telephony.

Uittreksel

Telefonie het ansienlik ontwikkel oor die jare, maar die basiese ouditiewe model waarin die klank van alle verskillende bronne bymekaar gemeng word na een enkelouditoriese stroom het nie verander sedert die eerste telefoon gebou is nie. Enkelouditoriese klank is baie moeilik om te volg in 'n meervoudigebron situasie, soos byvoorbeeld in 'n konferensie oproep.
Klank met oorsprong by 'n sekere punt in die ruimte sal 'n effens anderse pad na elke oor volg. Selfs is ons nie aktief bewus hiervan nie, verwerk ons brein hierdie ruimtelike aanduidinge om ons te help om klanke in die ruimte te vind. Dit is hierdie ruimtelike inligting wat ons toelaat om ons aandag te vestig en te luister na 'n enkele spreker in 'n omgewing waar baie verskillende bronne terselfdertyd aktief mag wees, 'n verskynsel wat bekend staan as die "skemerkelkiepartytjieeffek". Dit is moontlik om hierdie ruimtelike leidrade na 'n klank te reproduseer met behulp van hoofverwandeoordragfunksies (HRTFs) en om daardeur 'n luisteraar gelokaliseerde klank te laat ervaar, selfs wanneer die klank deur middel van oorfone gespeel word.
In hierdie tesis word ruimtelike klank geïmplementeer in 'n telefonieprogram, sowel as in 'n virtuelewêreld. Eksperimente is uitgevoer wat getoon het dat ruimtelike klank die verstaanbaarheid van spraak in 'n meerderebronomgewing verhoog en help met aktiewe spreker identifikasie. Hulpbrongebruiks metings toon aan dat hierdie voordele egter nie sonder 'n koste kom nie. Ter afsluiting, dit is bewys dat ruimtelike klank 'n verbetering tewees gebring het oor die enkelouditorieseklankmodel wat tradisioneel in telefonie gebruik het.

Acknowledgements

I would like to thank the following:

- My supervisors, Dr Gert-Jan van Rooyen and Dr Herman Engelbrecht, for their academic support and guidance.
- The open source community, for the development of some of the software used in my research.
- My family and friends, for their support and encouragement.

Contents

List of Figures

List of Tables

Listings

Chapter 1

Introduction

> One day every major city in
> America will have a telephone.
>
> Alexander Graham Bell

Telecommunication has certainly come far from that era, with South Africa having 92.2 cellular subscriptions for every 100 inhabitants in 2008 [128]. Although even when considering highly advanced devices such as the Apple iPhone, the fundamental audio model behind our modern voice-based telecommunication devices, of just mixing monaural audio signals together, has not changed since Bell was granted the first telephone patent in 1876 [48].

1.1 Motivation

Humans are generally social creatures, desiring interaction and communication with other individuals. Technology exists to augment our lives and one way in which it does this is by facilitating communication and interaction with others.

The Internet, through giving each user a virtual presence, is not impeded by the physical boundaries that normally limit human interaction. It is this virtual presence that makes social networks as an example so popular, and as a result, a highly profitable business [42]. Social networks may augment the way in which we stay in contact with others, but the basic communications model is no different from traditional means such as email and cellphone text messages. Telephony is similar, the functionality of our communication devices continues to improve, but the audio model remains unchanged, multiple audio streams are mixed together into a single monaural stream that is presented to the user. A new communications model is needed, one that is closer to how we communicate when in close physical proximity, while still giving users a virtual presence. Possessing two ears means that we have stereophonic hearing, which is wasted when we realise that our telephony systems are based upon a monaural audio model. There is clearly a need for a telephony system that makes use of stereophonic audio in some manner.

Sound is a vibration that propagates through a medium – be it a gas, liquid or solid [67]. Physically sound exists in a four-dimensional domain – each point in both time and space

has an intensity value assigned to it. Sound originating from a specific point in space will travel a slightly different path to each ear. Even though we are not consciously aware of it, our brain processes these spatial cues to help us to locate sounds in space [35]. It is this spatial information that allows us to focus our attention and listen to a single speaker in an environment where many different sources may be active at the same time; this is known as the "cocktail party effect" [41]. It is possible to reproduce these spatial cues in a sound recording using techniques such as the HRTF [55] and binaural recording [76, 59, 108, 25, 23] to allow a listener to experience localised audio, even when sound is reproduced through a headset.

Although the use of spatial audio is in relatively widespread use in the entertainment sector for example, surround sound in movies and computer games, it is not a common feature in telephony. Modern telephony generally makes use of the basic monaural audio model of just mixing the audio streams of different participants in a conference call together, thus discarding the spatial information [93]. Making use of a person's ability to separate sound based on perceived spatial location allows one to better communicate information than possible with traditional means [46]. Knowing who is currently speaking in a conference call can be important for a user. During a contract negotiation or job interview one would answer a questions differently depending on who is asking them, for example, one would answer a question asked by the lead developer differently than one asked by a member of human resources. People who do not regularly participate in conference calls might need assistance in identifying the active speaker. Although methods exist to determine the active speaker in a conference call from the data packets [61] and visually display this to the user, any method using monaural audio will not aid in situations where more than one speaker is active. Multiple microphone beamforming techniques can amplify the voice of the active speaker [38], but only work when the speakers are in the same locale and would not work well with multiple active speakers. In this work, we detail the design of a Voice over Internet Protocol (VoIP) system that utilises spatial audio and conduct experiments to demonstrate its effectiveness in assisting a user in speaker identification and to improve speech intelligibility.

1.2 Objectives

The broad purpose of the research is an investigation into possible uses of spatial audio in modern forms of electronic communication. The following are the primary objectives of the research:

- Do a background study on the psychoacoustics of spatial hearing.
- Study the mechanism by which spatial audio can benefit electronic communication.
- Study the existing means by which spatial audio is used in entertainment and communication.

- Develop and implement a prototype spatial audio communications system.
- Perform human-subject experiments to determine the validity of the hypotheses.
- Evaluate spatial audio implementation.

1.3 System Specifications

Potential benefits aside, any practical application will need to work within the bounds of certain constraints and limitations for it to be adopted by the public. A system with massive bandwidth or processing power requirements will not likely be successful.

1.3.1 Cost

A more expensive system is typically more difficult to sell to the public. Spatialisation that can be completely implemented in software, like HRTF and stereo panning, would have a great advantage here over a technology such as wave field synthesis, or even surround sound, due to the hardware costs involved.

1.3.2 Network Bandwidth

While bandwidth and broadband penetration levels are high in developed nations [29], developing nations like South Africa, the context for this project, are still severely lacking. In 2008, South Africa had a broadband penetration of only 10.5%. This is in part due to the high cost of broadband in South Africa which is 286% more expensive than a comparable offering in Egypt [28].

A VoIP call using the G.711 Pulse-Code Modulation (PCM) codec with a bit rate of 64 kbps requires a bandwidth of approximately 87.2 kbps per channel, for a total of 174 kbps for each call. With mobile broadband from Vodacom on a pay-as-you-use option of R2/MB [30] this translates to R153.28 per hour of VoIP usage. Therefore for spatial audio voice communications to succeed in such a climate it is absolutely necessary that additional bandwidth requirements are kept to a minimum.

1.3.3 Processing Power

Processing power considerations are important if the application is to be ported to a device with limited resources, such as a cellular phone. If this is to be done, some form of server-based approach will need to be taken.

1.3.4 Architecture

The system needs to be able to fit into existing VoIP architectures in order to be useful. Requiring a completely different architecture would greatly inhibit mass adoption of the

system. It would be beneficial to ensure that the system is compatible with a common VoIP protocol such as Session Initiation Protocol (SIP) for example.

1.4 Overview

A discussion of the literature that was studied for the purposes of this research is given in Chapter 2 on page 5.

Some practical applications of spatial audio are given in Chapter 3 on page 30, with more detail being given about a speech-based chat room and a more ambitious auditory environment.

Some theoretical work was required before beginning the design of the prototype application and this is given in Chapter 4 on page 36.

Chapter 5 on page 50 details the design and implementation of a proof-of-concept spatial audio VoIP application. The application is then intergrated into the virtual world Second Life.

The benefit and potential cost of spatial audio is evaluated by a series of practical experiments performed in Chapter 6 on page 71.

Finally, the research is concluded in Chapter 7 on page 106.

Chapter 2

Background

This chapter gives a summary of the literature necessary for successful completion of the project. The sound localisation ability of the auditory system gives us spatial hearing and is briefly discussed in Section 2.1 on page 5. Section 2.2 on page 6 gives an overview of past spatial hearing research and is followed by Section 2.3 on page 9 going into more detail about the psychoacoustics behind spatial hearing. Techniques by which spatial audio can be generated and presented to listeners are discussed in Section 2.4 on page 11. Some existing implementations of spatial audio are reviewed in Section 2.5 on page 16. The choice of audio transmission architecture is a large factor in the design of a telephony application and is discussed in Section 2.6 on page 17. The possible effect that bandlimiting the incoming audio signal can have on the integrity of spatial audio is discussed in Section 2.7 on page 19. Section 2.8 on page 22 gives an overview of some acoustic theory that will be required for the project. A further proof-of-concept application that was considered in Section 2.9 on page 27 is the integration of spatial audio in a virtual world environment. For this virtual world implementation, the avatar rotations will need to be converted into an Euler sequence using the quaternion transformations given in Section 2.10 on page 28. At summary of the chapter is given in Section 2.11 on page 29.

2.1 Sound Localisation

The ability of a listener to determine the range and direction of a sound is known as sound localisation. Sound localisation is the basis of the mechanism by which spatial hearing works.

If a sound source is not directly to the front of a listener then the sound will follow a different path to each ear, with one path being longer. The sound travelling to the further ear will be delayed more than the sound travelling to the closer ear resulting in a Interaural Time Difference (ITD), which can range between 0 μs for a source on the median plane and 650 μs for a source directly to one side [35, 47]. This longer propagation distance and shadowing by the head and torso will also attenuate the sound travelling to the further ear more, causing a Interaural Level Difference (ILD). The brain uses these binaural cues to locate sounds in space. Diffraction around the head reduces shadowing for low frequencies,

making the ILD frequency dependent which provides the brain with more localisation cues. The spectral content of the sound reaching the eardrum is further changed by resonance and reflections due to the shape of the pinna (the visible part of the outer ear) depending on the angle of incidence [119]. Localisation accuracy is dependent on the spectral content of the sound source, which means that a sinusoid is much more difficult to locate than a broadband source such as noise [53]. The combination of the ITD, ILD and acoustic filtering effect caused by the pinnae, head and torso is what gives us our ability to locate sounds in space [89, 67].

2.2 Spatial Hearing Research

In 1953 Cherry first formulated the "cocktail party" effect as, "how do we recognise what one person is saying when others are speaking at the same time?" [58]. Research by Cherry showed that listeners are able to focus on sounds presented to one ear and ignore sounds presented to the other. Since then this phenomenon has been the subject of many studies [54]. Results from a few of these studies will be discussed below.

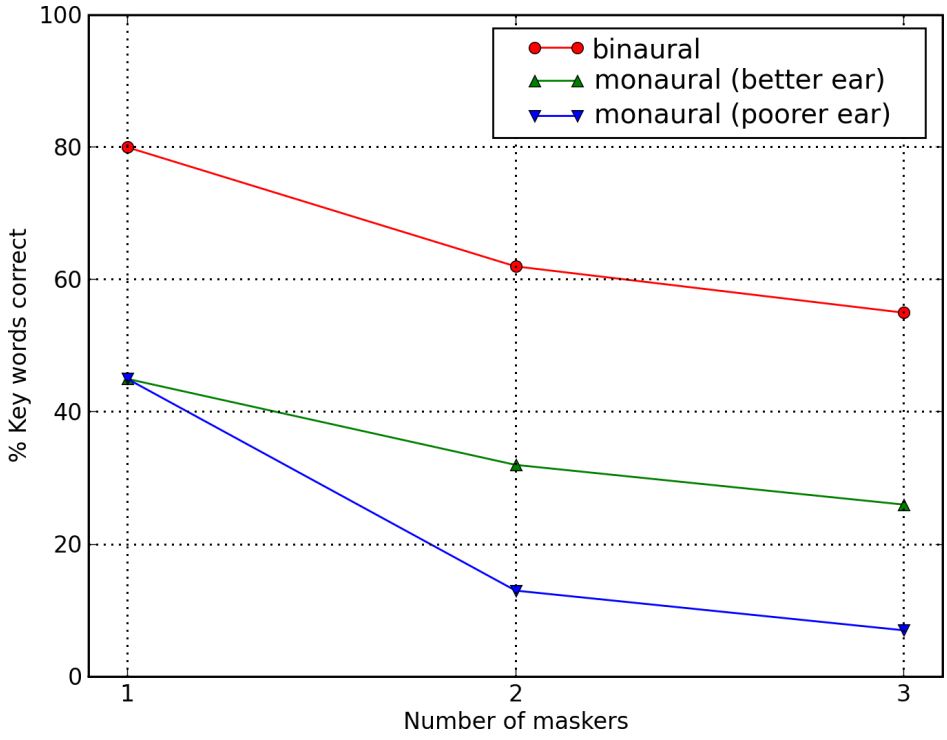
2.2.1 Hearing in Multiple Source Environments

Research has shown that spatial hearing is important for understanding a speaker in a multiple speaker environment because a listener can pay attention to sound originating from a specific direction while ignoring sound from all others [122]. The human hearing system easily suppress audio coming from a specific interfering azimuth while concentrating on another [114].

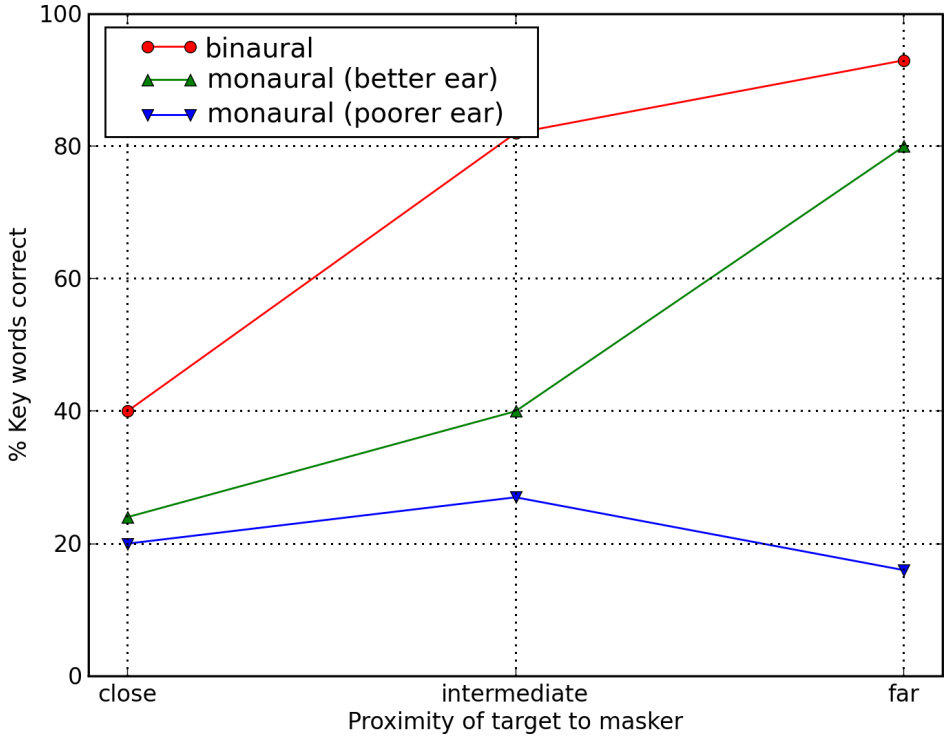
Hawley *et al.* performed an experiment with twelve subjects measuring the intelligibility rates of speech in the presence of masking speech presented binaurally and monaurally to both the subject's better and poorer ear, as a function of the number of maskers and positional separation between target and maskers [79]. They found higher performance with binaural presentation, as opposed to either monaural presentation method and that the binaural presentation offers a greater increase in intelligibility when increasing separation. A summary of their results is shown in Figure 2.1.

A study by Drullman and Bronkhorst shows an increase in speech intelligibility of three-dimensional audio over monaural or binaural audio in a situation with multiple active speakers [66]. The binaural presentation mode divided the speakers between the subject's two ears. The intelligibility rates for word and sentence targets is shown in Figure 2.2. It can clearly be seen that using HRTFs greatly increases a listener's ability to follow a target speaker in the presence of maskers. Generalised HRTFs are found to offer similar benefits as individualised ones.

If a listener has prior information on the expected spatial position of the target then performance increases [95]. This could be explained by the beamforming ability of the auditory system that focuses attention to a particular azimuth.

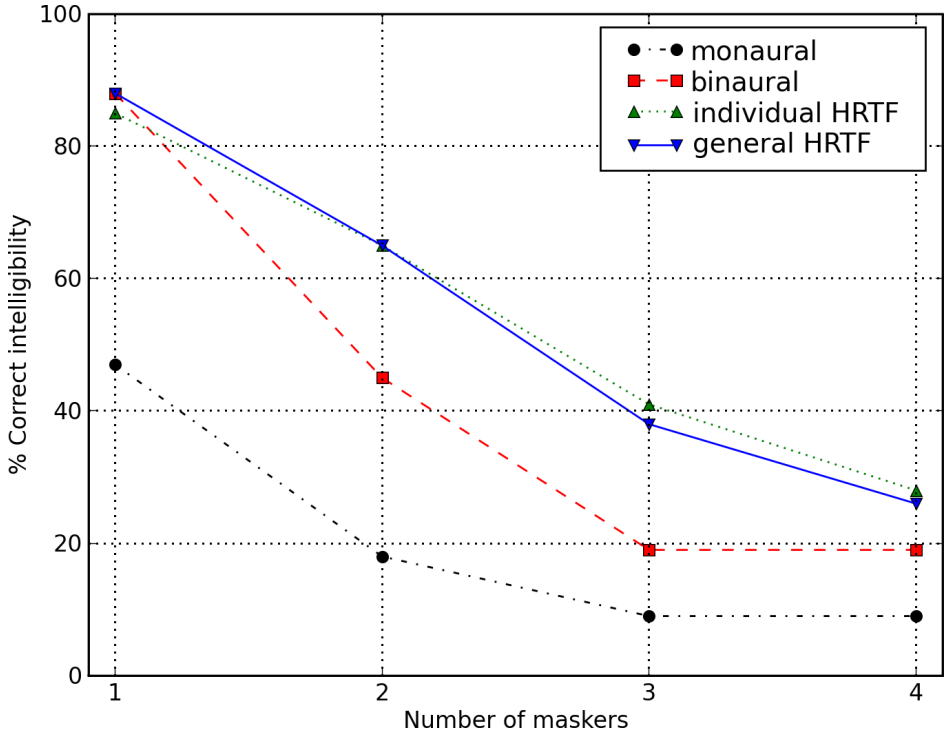


(a)

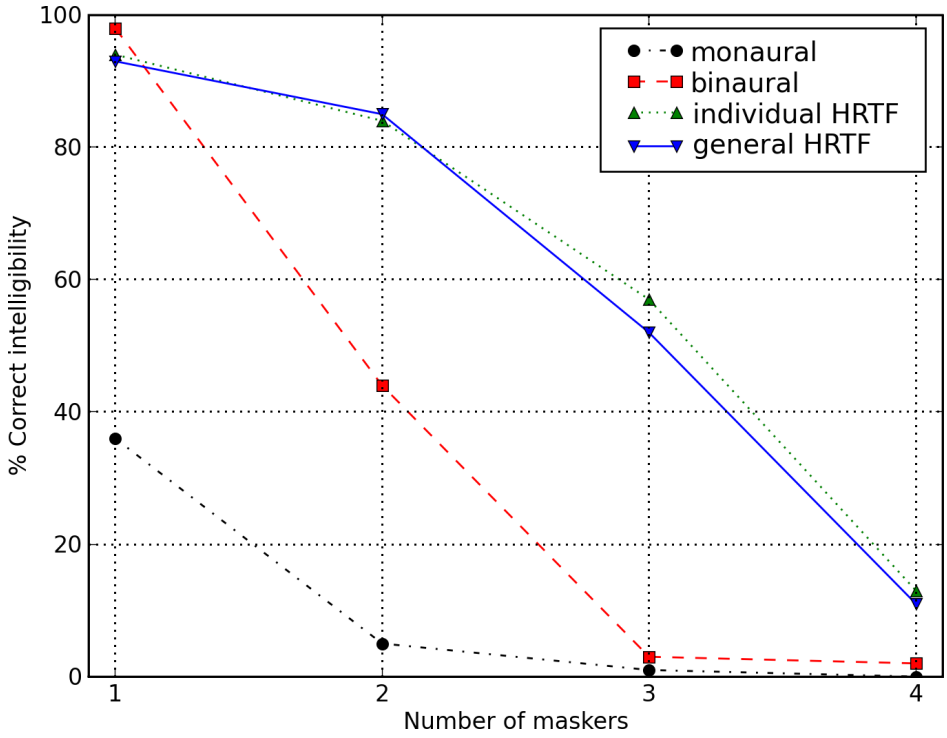


(b)

Figure 2.1: *Intelligibility as a function of masker number and configuration (data from [79]).*



(a) Words



(b) Sentences

Figure 2.2: *Intelligibility of multiple speaker speech (reproduced from [66]).*

2.2.2 Hearing in Noise

Patients with normal hearing in one ear and impaired hearing in the other, a condition known as unilateral hearing loss or single-sided deafness, have difficulty understanding speech in the presence of background noise [120]. Somewhat of a parallel can be drawn between single-sided deafness and the current monaural audio model used in modern voice telecommunication, neither provides any spatial separation between sound sources which can lower intelligibility in situations with more than one active sound source.

2.2.3 Locating Sound Sources

A flight simulator experiment conducted by Veltman *et al.* [129] showed that three-dimensional audio support increased performance of tasks that required information from a head-down display (HDD).

2.3 Psychoacoustics

Psychoacoustics is the study of the subjective perception of sound, while acoustics is the study of the physical properties of sound. A psychoacoustic study is important for research on spatial audio as it can further illuminate the mechanism in which spatial hearing works. Psychoacoustics can be used to play "tricks" on the auditory system and can create fake spatial separation between different sounds.

The human brain is easily fooled by "auditory illusions" and the perception of sound is highly subjective. If tones at 1000 Hz, 1200 Hz and 1400 Hz are heard together, then a pitch of 200 Hz is perceived to be heard because the brain misinterprets the tones as the 5th, 6th and 7th harmonics of a fundamental of 200 Hz [67]. Another auditory illusion is the Shepard tone [121], a series of ascending tones separated by octaves. The sound is perceived to continually rise in pitch indefinitely. Lossy audio compression codecs like MP3 use psychoacoustics with a model of the human auditory system to determine which parts of the spectrum will be masked and then to let quantisation noise rise in those segments, significantly reducing the data rate [67].

2.3.1 Simulating Source Direction via the Precedence Effect

The precedence effect is an auditory phenomenon that occurs when the same sound is heard from two different directions, with a slight time delay between them, which causes a phantom sound image to be heard close to the earlier source [119]. Sounds arriving within 50 ms of each other tend to be perceptually fused together and are not heard as separate sounds.

Freyman *et al.* performed an experiment to determine how important the type of spatial separation is on our ability to locate sounds in space [71]. They used the precedence effect to give the impression that the target source and masker were separated in space. The

sources were presented to a listener using two loadspeakers. A relative time delay of 4 ms between the signals presented to the two loudspeakers shifted the perceived location of the sound sources so that both either appeared to come from the front or the target from the front and the masker from the right. This perceptual spatial separation was compared to conventional spatial separation, where the target came from the front and the masker from the right for both a speech-spectrum noise interferer and a speech interferer. In the case of noise interference, the conventional spatial separation provided increased speech intelligibility whereas the perceptual separation provided a negligible advantage. Both approaches provided an increase in speech intelligibility for the speech interference case, although the conventional approach afforded a greater advantage than the perceptual one. From these results it is clear that merely exploiting the precedence effect to give the impression of spatial separation is not as effective as true spatial separation of sound sources and highlights the need for spatialisation algorithms that go further than merely introducing a time difference between the loudspeakers.

2.3.2 Type of Masking

When speech is in the presence of background noise, two types of masking occur—"energetic" and "informational" [39]. Energetic or peripheral masking occurs because all or part of the masker energy falling in the same frequency bands as that of the target signal, the overlapping of the masker and signal spectra results in decreased performance from the auditory system, specifically the cochlea which is thought to function as a series of auditory filters, the frequency to place mapping of which is described by the Greenwood function [74]. Informational masking occurs in the absence of this spectral overlap and is thought to result from the fusing of the target signal with the masker or due to uncertainty in the stimulus. Arbogast *et al.* [39] conducted an investigation into the effect of spatial separation of sources on the different types of masking. The target signal was at 0° azimuth and the masker was either 90° to the right or at the same location as the target. Sentences were filtered into 15 frequency bands by a modified cochlear-implant simulation program and pure tones at the centre frequencies of each band were modulated by the envelope of each band to create a preprocessed signal sentence. The target signal was created by summing eight of these frequency bands, randomly selected. A different-band sentence masker, which would result in primarily informational masking, was generated by summing six of the remaining seven frequency bands, randomly selected. A different-band noise masker was generated by convolving the different-band sentence masker with Gaussian noise. A same-band noise masker was generated by convolving the sum of the same eight bands used for the target signal with Gaussian noise. Both noise maskers had no intelligibility in speech and contributed little to informational masking. The advantage due to spatial separation for the different-band sentence masker (primarily informational masking) averaged 18 dB, the different-band noise masker (minimal energetic and informational masking) 4 dB and the same-band noise masker (mainly energetic masking) 7 dB. The main conclusion that Arbogast *et al.* drew

from this investigation is that spatial separation of sources leads to a greater advantage for informational masking than for energetic masking. We can conclude from this that spatial separation would be of a great benefit to persons in conference calls wanting to follow the speech of multiple participants, a situation resulting in mainly informational masking.

2.3.3 Effect of Spatial Configuration

The effect of positional configuration always makes a larger difference than adding or removing a single competing speaker [79]. This means that the loss in speech intelligibility resulting from adding another speaker to a conference call can be made up by changing the positional configuration to one more favourable for the number of sources. The intelligibility of a conference call with any number of speakers can therefore be maximised by taking advantage of this.

2.4 Spatial Audio Rendering Techniques

We define spatialisation as giving an auditory source actual or perceived direction through either physical or psychoacoustic measures. A number of ways exist to present spatial or audio that has perceived direction to a listener. Some use multiple sources to create positional audio and other use psychoacoustics to "trick" the listener into believing that the sound comes from the apparent direction. A few of these techniques will be discussed in this section.

2.4.1 Surround Sound

The surround sound technique uses multiple loudspeakers to encircle the listener [83, 119]. Sound sources are panned between the different loudspeakers to give it direction. Many different arrangements exist, for example, 5.1 surround sound which has a left, centre, right, left surround, right surround and low frequency effects channels. The sound images are only perceived correctly in a small area; this limited "sweet spot" means that the effect is not accurate for a large or moving audience.

2.4.2 Binaural Recording

Binaural recording is a method in which a high-fidelity microphone is placed inside each ear canal of a human subject or dummy model [76, 59, 108, 25, 23]. The resulting stereo recording then incorporates the binaural filtering effects of the subject's pinnae, head and torso and reproduces the sound pressures exactly as they are at the subject's eardrums. If the recording is played back to a listener through stereo headphones they will experience the recording as if they heard it in person. Localisation is influenced by the degree of similarity between the anthropometric measurements of the listener and subject used to make the recording.

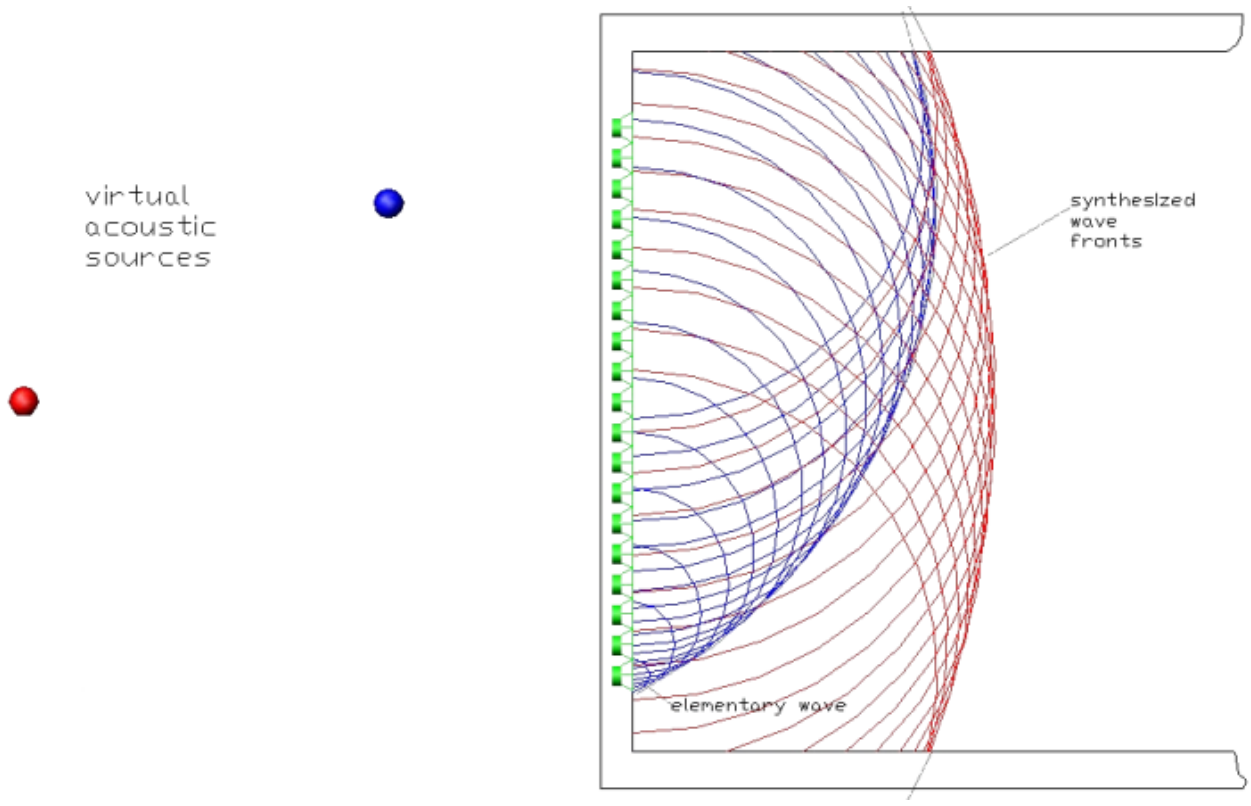


Figure 2.3: *Wave field synthesis (reproduced from [113]).*

Binaural recording complicates the recording of samples and the technique cannot be used to create localised audio after the fact. Binaural recording is best suited for playback over headphones, they do not sound very good when played on loudspeakers without additional signal processing [119].

2.4.3 Wave Field Synthesis

Wave field synthesis is based on the Huygens' principle [49], which states that a wave front can be considered the superposition of elementary spherical waves, and is quantified by the Kirchhoff-Helmholtz integral [136]. Wave field synthesis creates the wave field of a virtual acoustic source by superposition of elementary waves from a large loudspeaker array [50, 126, 113] as shown in Figure 2.3. Sound reproduction is excellent, but this method is expensive to implement due to the high number of required loudspeakers and careful calibration, which is dependent on listening room acoustics.

2.4.4 HRTFs

As discussed in Section 2.1 on page 5, the spectral filtering resulting from the different paths that a sound source from an arbitrary point in space travels to reach each ear provide cues that aid a listener in locating the sound source. This spectral filtering can be expressed by

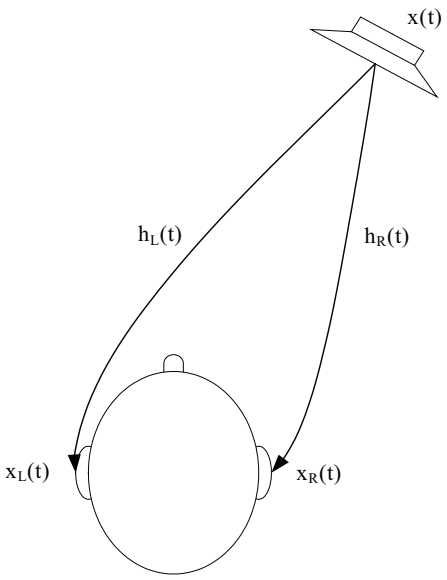


Figure 2.4: *HRTFs express the spectral filtering due to the different path travelled by the sound from a source to each ear.*

HRTFs as shown in Figure 2.4. Strictly speaking, the term HRTF refers to the frequency-domain function and Head-Related Impulse Response (HRIR) the time-domain function thereof, but most literature will speak just of an HRTF, the domain implied in the context. We will also use the term HRTF to refer to both the time-domain and frequency-domain functions. HRTFs can be measured, and such databases [7, 36, 72, 22] can be used to give sound the perception of direction by convolving a monaural audio source $x(t)$ with an HRTF pair giving outputs

$$x_L(t) = x(t) \star h_L(t) \tag{2.1}$$

and

$$x_R(t) = x(t) \star h_R(t), \tag{2.2}$$

with $h_L(t)$ and $h_R(t)$ being the HRTFs for the left and right ear respectively. Output signals $x_L(t)$ and $x_R(t)$ are respectively the sound heard by the left and right ear. The HRTFs of every person are unique and the brain undergoes a constant calibration process to ensure accurate sound localisation [82]. Research has shown that generalised HRTFs provide localisation comparable to that of free-field sources for approximately 75% of subjects [135], with the remainder having difficulty discriminating elevation. Localisation accuracy is expected to improve with training or when individualising the HRTFs [137, 138].

Figure 2.5 shows the time truncated HRTF pair for an azimuth of 30° and clearly shows that the initial wavefront reaches the right ear first, due to it being closer to the source than the left ear. The greater distance also means that the audio heard at the left ear experiences greater distance attenuation. Figure 2.6 shows a mesh plot of HRTFs for the right ear on the horizontal plane.

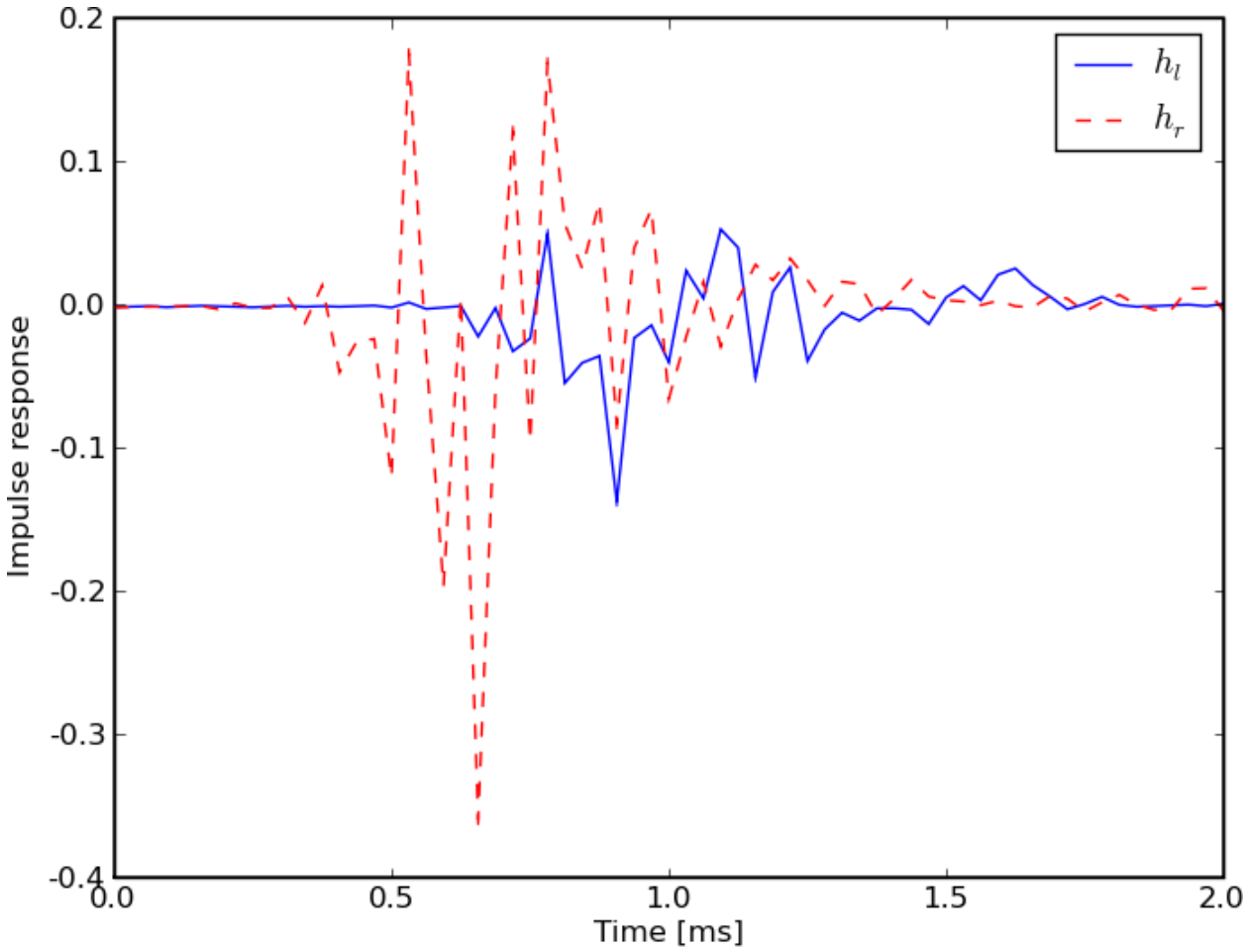


Figure 2.5: *HRTF pair for azimuth*=30°.

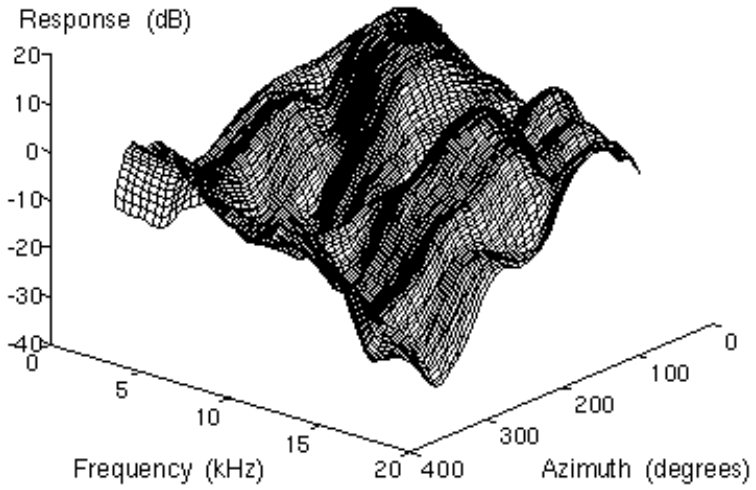


Figure 2.6: *Right ear HRTF on the horizontal plane (reproduced from [1]).*

Table 2.1: *Measurement points for the KEMAR HRTF database (reproduced from [72]).*

Elevation	Azimuth increment	Points per elevation
-40°	6.43°	56
-30°	6.00°	60
-20°	5.00°	72
-10°	5.00°	72
0°	5.00°	72
10°	5.00°	72
20°	5.00°	72
30°	6.00°	60
40°	6.43°	56
50°	8.00°	45
60°	10.00°	36
70°	15.00°	24
80°	30.00°	12
90°	360.00°	1

Virtual surround sound can be created with two or three loudspeakers by using HRTFs to create "virtual" surround loudspeakers [119]. The end result is reasonable at best, with most systems providing poor reproduction due to heavy timbral modification. Timbre is the "shape" of a sound, a dumping ground for any qualities not normally attributed to loudness or pitch.

Sources of HRTFs

Measuring HRTFs is a complex task [7, 72] and it is simpler to just make use of one of many freely available HRTF databases. A few of these will now be discussed.

KEMAR HRTF Database

The Knowles Electronic Manikin for Acoustic Research (KEMAR) database is a set of HRTF measurements from a KEMAR dummy head done by Bill Gardner and Keith Martin at the MIT Media Laboratory [72]. The HRIRs are 512 samples long, sampled at a temporal rate of 44.1 kHz and quantised to 16 bits. Table 2.1 shows the measurement points.

Listen HRTF Database

The Listen database is the result of HRIR measuring sessions done by Ircam and AKG as a part of the greater Listen project [7]. The database has measurements for 51 subjects. The measured impulse responses are 8192 points long, sampled at 44.1 kHz and quantised to 24

Table 2.2: *Measurement points for the Listen HRTF database (reproduced from [7]).*

Elevation	Azimuth increment	Points per elevation
-45°	15°	24
-30°	15°	24
-15°	15°	24
0°	15°	24
15°	15°	24
30°	15°	24
45°	15°	24
60°	30°	12
75°	60°	6
90°	360°	1

bits. Equalisation and post processing, which includes removal of the propagation delay, is done on the measured impulse responses giving HRIRs that are 512 points long. The HRIRs are not measured in equal increments at all elevations; Table 2.2 shows the measurement points. On the horizontal plane the HRIRs are measured in 15° increments. The azimuth is measured clockwise, with 0° being the direction that the subject is facing and 90° directly to the right. The loudspeakers are positioned 1.95 m from the subject. The elevation is measured with 0° being in the direction that subject is facing and 90° directly above.

CIPIC HRTF Database

The CIPIC HRTF is a public domain database of HRTF measurements done at the University of California, Davis CIPIC (Centre for Image Processing and Integrated Computing) Interface Laboratory [36]. The database has measurements for 45 subjects, 27 male, 16 female and the KEMAR with large and small pinnae. The HRIRs are 200 samples long, sampled at a temporal rate of 44.1 kHz and quantised to 16 bits. Elevations were uniformly sampled in 5.625° increments from $-45°$ to 230.625°. Azimuths were sampled at $\pm 80°$, $\pm 65°$, $\pm 55°$ and in 5° increments from $-45°$ to 45°. The azimuths were chosen to obtain approximately uniform density on the sphere, leading to spatial sampling at 1250 points.

2.5 Implementations of Spatial Audio

Encountering spatial audio is common in the entertainment sector with surround sound loudspeaker systems prevalent in cinemas and home theatre set-ups as well as in computer and console games [44]. Reproduction of surround sound on headphones is done using psychoacoustics and HRTFs [21, 26].

DiamondWare offers a positional audio solution for software developers [3], but this is a

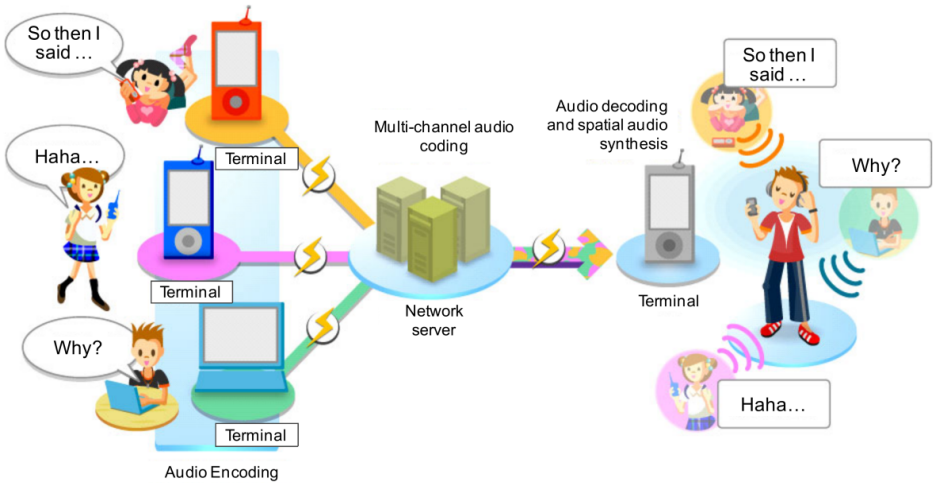


Figure 2.7: *NTT DOCOMO Spatial Audio Transmission (reproduced from [112]).*

proprietary service and they do not provide technical information on how this is implemented. DiamondWare provides the positional audio implementation found in the proprietary Second Life voice communications system developed by Vivox.

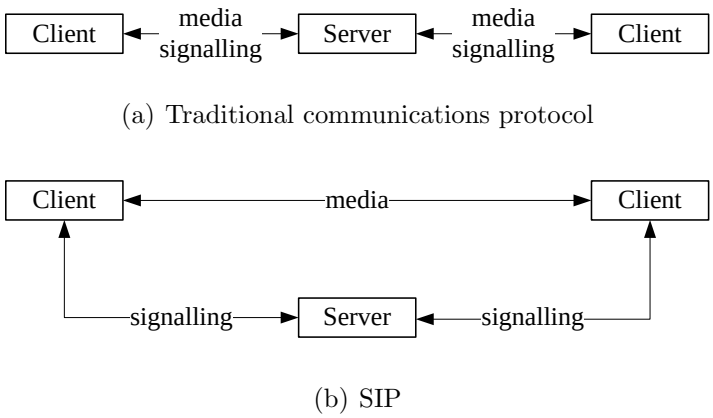
Japanese mobile phone operator NTT DOCOMO has recently developed spatial audio transmission technologies for use on their Pro HT-01A cellular handset, the Japanese equivalent of the HTC Touch Pro [112]. The technology, as shown in Figure 2.7, employs an approach that has the server only transmit the important auditory components of the speakers' voices to the client where the spatialisation is performed. This approach reduces the bandwidth and processing load placed on the clients, but requires a server which increases implementation costs.

Voiscape is a prototype spatial audio communications medium in development that attempts to create virtual auditory spaces for people to communicate in [92, 93]. The first prototype used a special sound card that implements HRTF spatialisation in a chip and the second features a 3D voice server for spatialisation. The system has distance-attenuated, directional audio. Evaluation of the system has been limited and experimentation is needed to determine the value of spatial audio in the context of the application being developed.

OnLive Traveler is a voice communications system that uses stereo panning and distance attenuation to create a virtual auditory environment [64]. Users navigate the environment through a controllable avatar. The audio processing is handled by a central server.

2.6 Telephony Architecture

The prototype application will need to be developed on a platform that is used by a large portion of the population and allows for processing of the raw audio signals. A software-based system such as VoIP is a much better solution for rapid development than something reliant on hardware like a Public Switched Telephone Network (PSTN) system.



(a) Traditional communications protocol

(b) SIP

Figure 2.8: *SIP provides separation in media and signalling.*

2.6.1 VoIP

VoIP is becoming increasingly common in modern telecommunication, in part due to its ease of deployment and lower cost compared to traditional PSTN systems [99, 40]. The United States Federal Communications Commission is currently in early preparation for switching the old PSTN in the United States to a system based entirely on VoIP [68]. If a call is made using a service like Skype that does not charge when calling within the VoIP network, then the physical distance between the end-points does not affect the cost of the call. A call made between two continents will cost the same as a call made within the same city and the only cost is the Internet bandwidth used [19]. Using VoIP reduces infrastructure costs as both data and voice communications utilise the same network [124].

2.6.2 SIP

A VoIP system can be implemented using a variety of protocols, each with its own advantages and disadvantages. SIP will be discussed in the remainder of this section.

SIP [118] is a signalling protocol for handling the creation and termination of multimedia communication sessions. SIP differs from other more traditional telecommunications protocols in that it does not require the media to go through the server that initiated the session, as shown in Figure 2.8. This separation of signalling and media and the fact that the server only establishes a connection and leaves the clients to do the media streaming on their own. This allows SIP to be used for novel applications not yet thought of when the protocol was designed, such as including video or positional information for constructing three-dimensional audio spaces. It also lowers costs as the server does not need to process or route the media. A SIP VoIP network can be connected to a PSTN using a gateway that maps the signalling and media transport protocols between the two domains [124], allowing SIP to coexist with traditional telephony networks.

Figure 2.9 shows an example of the functioning of a SIP session [124]. When a SIP user

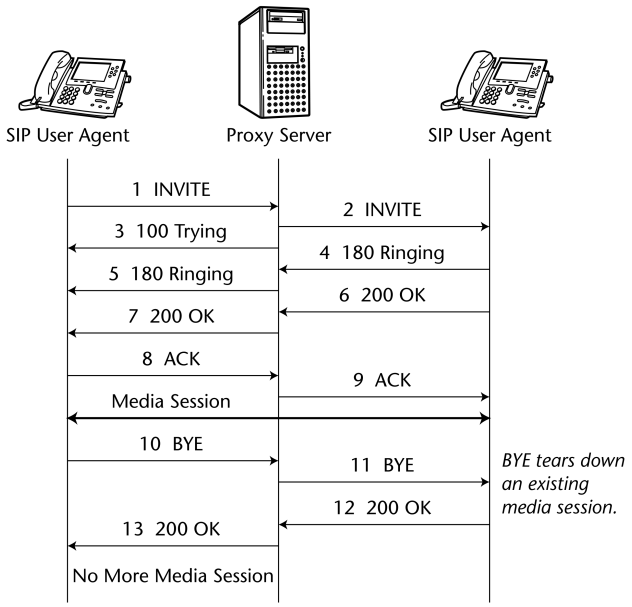


Figure 2.9: *Example SIP session (reproduced from [124]).*

agent wishes to communicate with another client they send a request via a SIP proxy server. The server serves to relay signalling between the two user agents. Once a session has been established, the media flows directly between the clients in a peer-to-peer fashion using the Real-time Transport Protocol (RTP). The protocol does not impose any restrictions on what type of media is allowed, be it voice or video, or on which codecs are used. Upon completion of communication the media session is torn down.

SIP calls can be initiated without a server if the SIP Uniform Resource Identifier (URI) of the remote client is known as shown in Figure 2.10, signalling and media is handled using a peer-to-peer approach. This approach is generally impractical as it necessitates knowing the Internet Protocol (IP) address of the client to be called.

While SIP does not route any media through a server, a server is still required for the initiation of sessions. Peer-to-peer SIP is an implementation of SIP using a peer-to-peer architecture to provide a distributed VoIP communications system [123]. Peer-to-peer systems are inherently scalable and, because of the lack of a single point of failure, reliable.

2.7 Effect of Bandwidth on Spatialisation

Any VoIP platform will enforce certain signal bandwidth constraints which could negatively impact spatial audio.

The bandwidth of the source audio signals will have a greater effect on spatial than monaural audio. In speech signals, the frequencies below 8 kHz are sufficient for accurate speech recognition, but information essential to localisation is contained in natural speech in the band above 8 kHz [51]. When the speech is bandlimited to 4 kHz, articulation

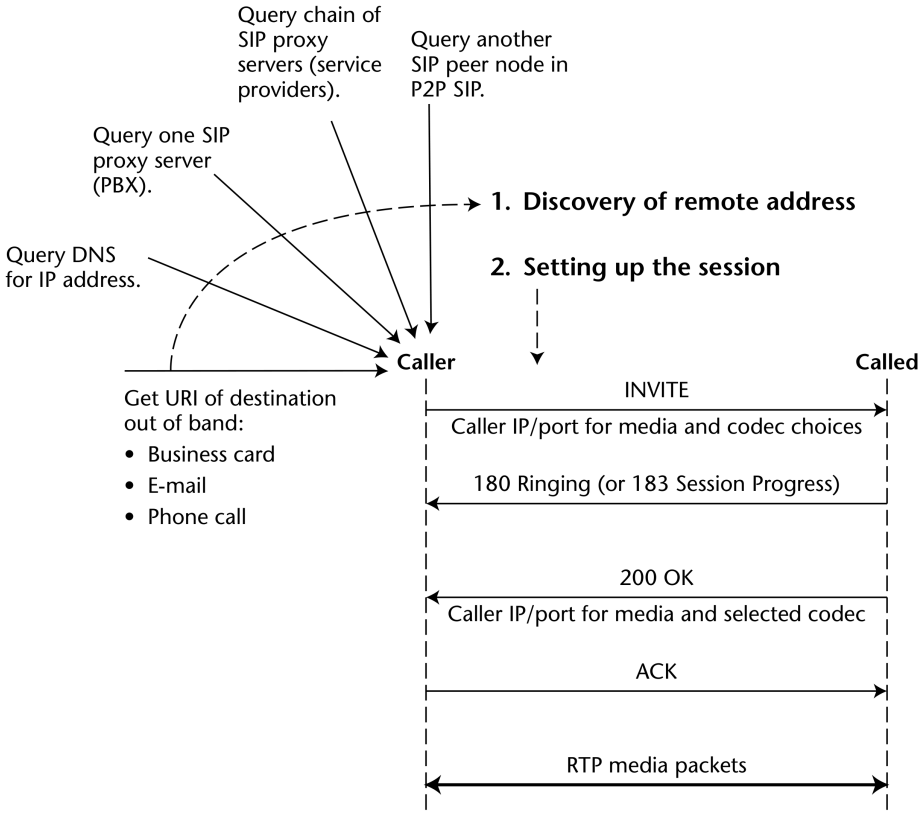


Figure 2.10: *Basic SIP session (reproduced from [124]).*

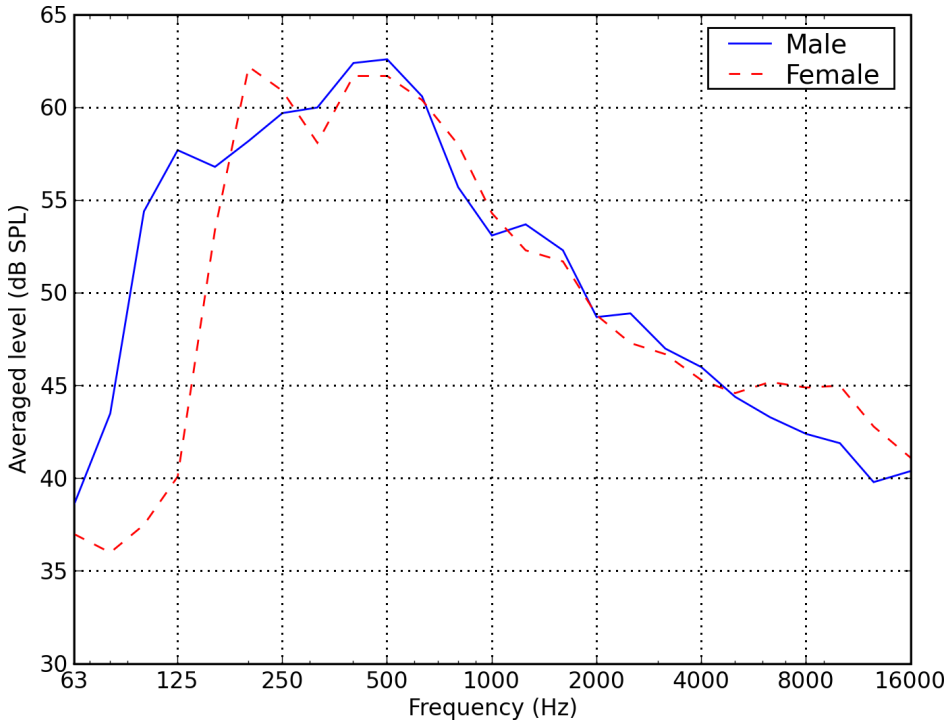


Figure 2.11: *Averaged speech spectra, normalised for 70 dB overall level (reproduced from data in [56]).*

begins to suffer, and consonants are especially effected [70, 80]. The spatialisation accuracy is affected by the bandwidth of the signal being spatialised; if the signal does not occupy the full the bandwidth of the HRTF, then only sections of the spectral shaping will be implemented. Results from a study on average long-term speech spectra by Byrne *et al.* [56], summarised in Figure 2.11, shows that there is speech energy located in the 8-16 kHz band (the study measured the spectra between 63 Hz and 16 kHz). As energy is contained in the band above 8 kHz, it would seem that keeping this band would benefit spatialisation. A Finite Impulse Response (FIR) filter performs spectral shaping [115] and cannot create new frequency components. Filtering a signal bandlimited to 8 kHz with a filter bandlimited to 16 kHz results in an output bandlimited to 8 kHz. King and Oldfield found that the minimum bandwidth for accurate elevation and front-back discrimination is 0-13 kHz or 1-16 kHz and that for high spatial resolution, a signal bandwidth of 0-16 kHz is necessary [96]. Best *et al.* compared localisation accuracy of a broadband corpus (300 Hz - 16 kHz) and the same corpus low-pass filtered at 8 kHz and found that the filtered sound signals degraded localisation performance [51]. From this it is clear that, if a spatial audio telecommunication system is to be successful, a speech codec and system with a sampling rate of 32 kHz will need to be used in order to achieve the 16 kHz signal bandwidth necessary for accurate localisation.

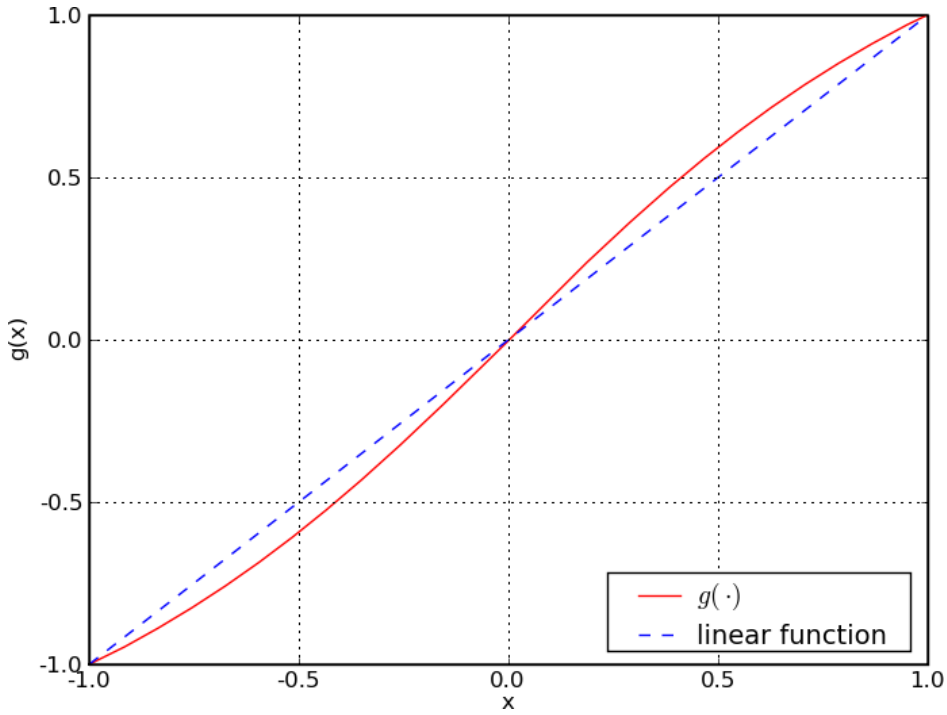


Figure 2.12: *Distortion function* $g(\cdot)$.

2.7.1 Bandwidth Expansion

The bandwidth of a signal with limited spectral content can be expanded by introducing harmonics into the signal, increasing the effectiveness of the spatialisation process [62].

The bandwidth of an audio signal $s(t)$ can be modified by passing it through a non-linear system [91], giving a distorted output signal

$$d(t) = g[s(t)], \tag{2.3}$$

where $g(\cdot)$ is some nonlinear function.

Let $s(t)$ be a 440 Hz sinusoidal signal and $g(s(t)) = tan^{-1}(s(t))$ as shown in Figure 2.12, the distorted signal is then as shown in Figure 2.13. Figure 2.14 shows $S(f)$ and $D(f)$, the frequency-domain representations of $s(t)$ and $d(t)$.

2.8 Acoustics

This section gives some underlying theory relating to acoustics that was required for the project.

2.8.1 Reverberation

Reverberation was considered but ultimately not used in this project because of the extreme computational cost and very accurate acoustics not being essential in a system where con-

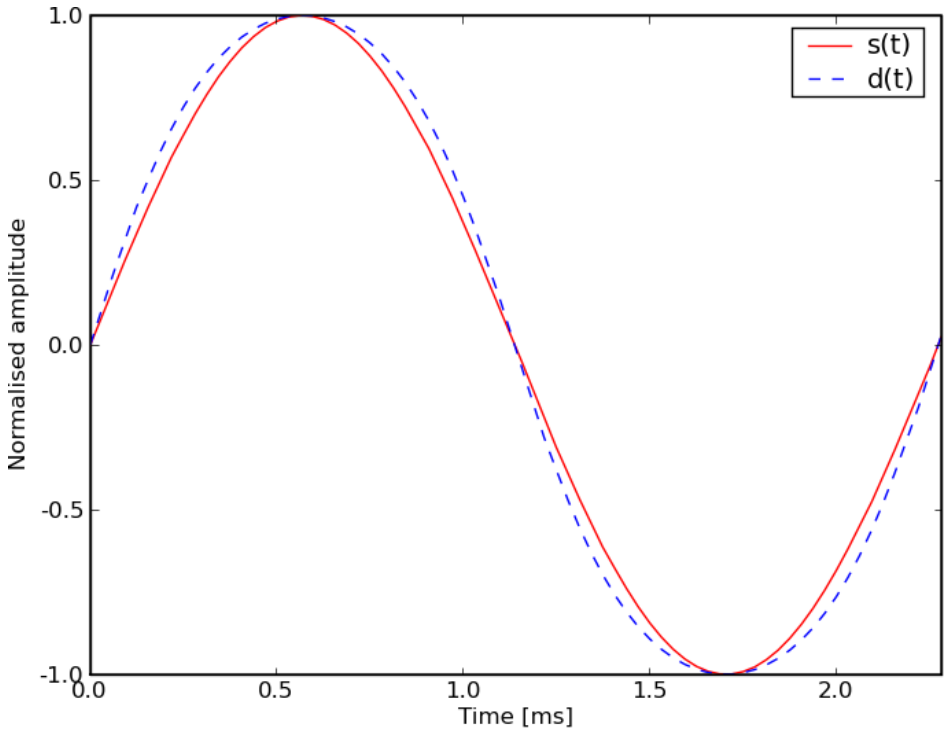


Figure 2.13: *Distorted signal, $d(t)$, compared to the undistorted signal, $s(t)$.*

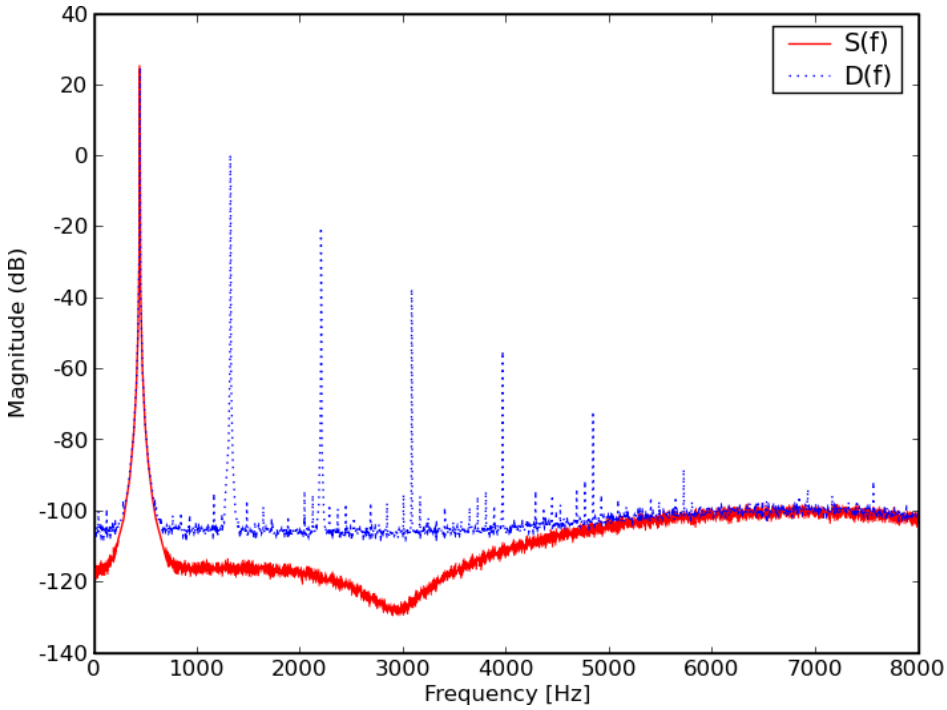


Figure 2.14: *Spectrum of a distorted signal, $D(f)$, compared to that of the undistorted signal, $S(f)$.*

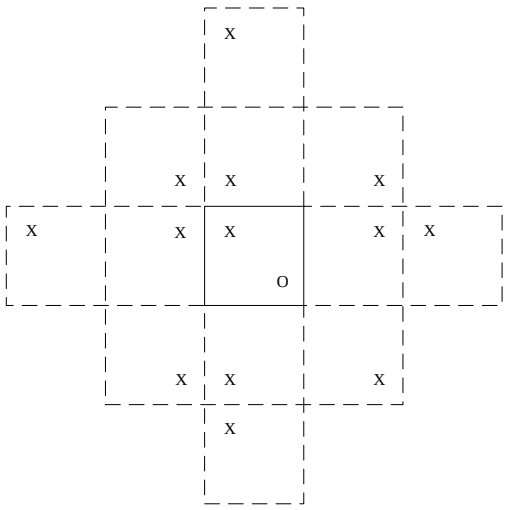


Figure 2.15: *Image method for calculating room reverberation. X represents the source, O the listener and the solid box the original room (reproduced from [37]).*

versation is the main focus. If more realism is required then reverberation will need to be implemented. A discussion on reverberation is given in this section because it would likely benefit future research work on spatial audio.

When a burst of sound is emitted in some environment, the intensity will slowly decay as it bounces around its environment. This is called reverberation [67] and depends on the acoustics of the environment. A simple measure of the nature and amount of reverberation that a room possesses is reverberation time, the time required for the sound to decay by 60 dB.

Reverberation can be simulated by convolving the audio stream with a room impulse response, which gives accurate modelling but is computationally expensive, or by using nested networks of allpass filter with feedback, which requires less processing power but produces artefacts and sounds "metallic" [110].

The image method is a model for calculating the room reverberation impulse response [37]. The general idea of the image method is to create a specular "mirror image" of the source in each surface of the room [47]. Higher order reflections are mirror images of mirror images. The images can be seen as virtual copies of the source. Vectors can be drawn from all sources, real and virtual, to the listener and the length of these vectors used to calculate the reverberation impulse response. Figure 2.15 shows a two-dimensional representation of the model up to second reflections, viewing the room from the top. The model does not take into account occlusion of sound by objects in the room.

The image method gives the room impulse response for rigid walls at time t, for a source at position $S = (x_s, y_s, z_s)$ and a listener at position $M = (x_m, y_m, z_m)$ as

$$p(t, S, M) = \sum_{p=0}^{7} \sum_{r=-\infty}^{\infty} \frac{\delta\left(t - \left|R_P + R_r\right| / c\right)}{4\pi \left|R_P + R_r\right|}, \tag{2.4}$$

where $c = 343m/s$ the speed of sound, $Rp = (x_s \pm x_m, y_s \pm y_m, z_s \pm z_m)$, $R_r = (2nL_x, 2lL_y, 2mL_z)$ where (n, l, m) is an integer vector triplet and (L_x, L_y, L_z) the room dimensions.

When the room walls are not rigid, the acoustic reflection coefficients of the walls need to be taken into account. The model given in Equation 2.4 assumes perfect reflection of each image. The room impulse response now becomes

$$p(t, S, M) = \sum_{p=0}^{7} \sum_{r=-\infty}^{\infty} \beta_{x1}^{|i-d|} \beta_{x2}^{|i|} \beta_{y1}^{|j-e|} \beta_{y2}^{|j|} \beta_{z1}^{|k-f|} \beta_{z2}^{|k|} \times \frac{\delta\left(t - \left|R_P + R_r\right| / c\right)}{4\pi \left|R_P + R_r\right|}, \tag{2.5}$$

where $R_p = (x_s - x_m + 2dx_m, y_s - y_m + 2ey_m, z_s - z_m + 2fz_m)$ is now expressed in terms of the vector triplet $p = (d, e, f)$ and the β terms are the reflection coefficients of the walls with β_{x1} being the top wall, β_{x2} the bottom wall, β_{y1} the left wall, β_{y2} the right wall, β_{z1} the floor and β_{z2} the ceiling of the room shown in Figure 2.15.

Reverberation is also used as a measure of distance, and artificial reverberation can be used to add a sense of space to an artificial acoustic scene [119]. As the distance between a source and listener increase, the ratio between reverberant and direct sound increases, with the reverberant radius of an environment being the distance at which the reverberant sound energy equals the direct sound energy [54]. For sound generated by a point source, the reverberation radius is

$$r_h = 0.1\sqrt{\frac{V}{\pi T}} \tag{2.6}$$

where V is the volume of the room (in m^3) and T the reverberation time of the room (in seconds). When the sound is not generated by a point source, the reverberation radius should be multiplied by the directivity of the source. The directivity of a source is the ratio of the intensity in the direction being considered to the average intensity of the source.

2.8.2 Plenacoustic Function

The minimum allowable angular interval in between adjacent HRTF pairs will need to be determined.

Much in the way that the Nyquist frequency specifies the minimum temporal sampling frequency that allows perfect reproduction, a similar spatial sampling frequency exists. This spatial sampling frequency would set the maximum spatial sampling internal that would allow for perfect reproduction of HRTFs for any position.

The plenacoustic function, named in reference to the plenoptic function, characterises the soundfield at any point in space. The plenoptic function is a seven-dimensional function $f(\theta, \phi, \lambda, t, V_x, V_y, V_z)$ that represents the intensity of light when looking in a direction (θ, ϕ), at wavelength λ, at time t and at location (V_x, V_y, V_z) [31]. The plenacoustic function $p(x, y, z, t)$ is a collection of impulse response and is defined as the sound pressure recorded at location (x, y, z) at time t, and once sampled allows us to reproduce the soundfield at any point [34]. Being continuous, the plenacoustic function needs to be sampled at a minimum temporal and spatial frequency if aliasing is to be avoided.

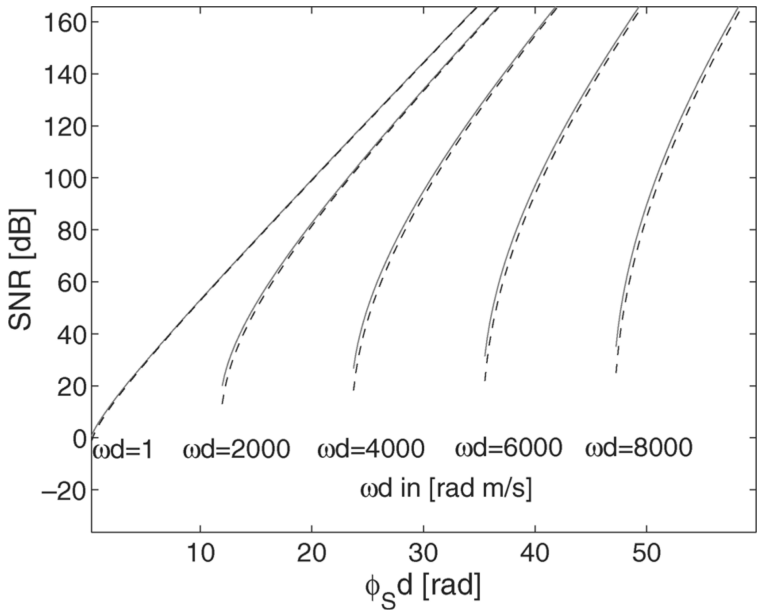


Figure 2.16: *Plenacoustic reconstruction for a given SNR (reproduced from [34]).*

The first case considered is that of the impulse responses along a line in a room. The relationship between temporal and spatial frequencies is given as

$$\omega_d = \frac{2\omega_t}{c}, \tag{2.7}$$

with c being the propagation speed of sound in air (343 m/s) and $\omega_t = \frac{2\pi}{\Delta t}$ and $\omega_d = \frac{2\pi}{\Delta d}$ being the temporal and spatial pulsations respectively with Δt being the temporal sampling period of the impulse responses and Δd the spatial sampling interval between the impulse responses [34].

The impulses responses of a rectangular room can be reconstructed for a maximum temporal frequency ω_0 at a given Signal-to-Noise Ratio (SNR) when the spatial sampling frequency $\phi_s = 2\pi/\Delta x$, where Δx is the spatial interval between sampling points, satisfies

$$\phi_s > \frac{2\omega_0}{c} + \epsilon\left(\mathrm{SNR}_0, \omega_0\right), \tag{2.8}$$

where $\epsilon\left(\mathrm{SNR}, \omega\right)$ is obtained from Figure 2.16.

For the case of HRTFs the angular sampling frequency needs to satisfy

$$l_{\theta_S} > 2\omega_{max}\frac{d}{2c} \approx 2\omega_{max}\frac{0.09}{c}, \tag{2.9}$$

with $l_{\theta_S} = \frac{2\pi}{\Delta\theta}$ being the angular sampling frequency, $\Delta\theta$ the angular spacing, ω_{max} the maximum temporal frequency [33] and the inter-aural distance (spacing between the two ears) being $d = 18$ cm [53]. This means that the minimum angular spacing allowed is

$$\Delta\theta_{min} = \frac{2c}{f_s d}, \tag{2.10}$$

A minimum angular spacing of 4.95° is therefore required for HRTF interpolation at a temporal sampling frequency of 44.1 kHz and 6.82° at 32 kHz.

2.8.3 HRTF Interpolation

Simulating moving sound sources requires HRTFs with a low spatial sampling interval for smooth movement between positions [47]. If the selected HRTF database does not provide this, then new HRTFs in between the given ones will need to be interpolated.

Begault states that HRTFs should ideally be interpolated in the frequency domain, but that time domain linear interpolation can be done as long as the ITD is interpolated separately [47]. Nishino *et al.* showed that although non-linear interpolation gives a better signal-to-deviation ratio than linear interpolation, there is no significant difference in subjective evaluation of HRTFs generated with either method [111]. Matsumoto *et al.* showed that arrival time correction improved interpolation accuracy with linear interpolation as long as HRTFs exist for enough azimuth values [106]. This agrees with plenacoustic theory stating that a minimum spatial sampling frequency exists for HRTF reconstruction and interpolation.

2.9 Virtual Worlds

Virtual worlds are a good example of an application in which we believe spatial audio will provide for a more immersive user experience. The remainder of this section will discuss some of the dynamics of virtual worlds.

On-line games are becoming increasingly common, with 78% of teenage and 50% of 18-32 year old Internet users playing games on-line [90]. In the second quarter of 2009, virtual world memberships are reported to have grown by 39% to an estimated 579 million [94]. Not all players crave the excitement of a game like Blizzard's World of Warcraft and some prefer a more relaxed, social gaming experience. Second Life is a on-line, three-dimensional virtual world developed by Linden Labs that allows users to interact with each other in a large world [16] and has millions of registered users and around 100,000 active users [134].

Second Life has a higher user engagement time than traditional social networks, with users spending an average of 100 minutes in-world per visit [103]. Second Life could even be seen as a rival to social network sites [102]. Monetising the product, an area in which traditional social networks are lagging, is something that the microtransaction model of virtual worlds like Second Life solves [103, 127]. Virtual worlds are a profitable business, with a revenue of $1 billion in 2008 and forecasted revenue of $17.3 billion in 2015 [127]. Microtransactions account for approximately 86% of this. Second Life boasts a bustling economy with more than $1 billion in transactions between users during the time frame between the second quarters of 2008 and 2009 [103].

Virtual worlds are not just for playing with: businesses are starting to use them to facilitate training, networking and real-time collaboration [131] due to lower costs involved compared to audio or web conferences [104]. Conferences in the physical world can be augmented by allowing those who could not attend in person to participate by means of a virtual world, providing a basis for mixed-reality meetings [65].

IBM is currently developing a real-time, three-dimensional collaboration environment, a service that the company calls Virtual Collaboration for Lotus Sametime or "Sametime 3D" [86, 73]. Sametime 3D is a service that IBM feels will replace instant messaging for remote business collaboration. Sametime 3D is to have a positional audio model that also has audio volume increase as avatars approach each other. The service is built using Second Life and OpenSim. The environment is to be more interactive than current user interfaces.

With more than 18 billion minutes of voice having been used between September 2009 and when the service launched in February 2007, Second Life is also starting to emerge as a major VoIP provider [103, 102, 45]. The 15 billion voice minutes that Second Life is forecast to handle in 2009 is still less than the 65 billion handled by VoIP giant Skype, but comparing the 700,000 active users of Second Life to the 42 million users of Skype shows a high participation in Second Life's VoIP service. One possible explanation is that a virtual world fosters dynamic conversations resulting in people conversing more freely than in an environment that amounts to an Internet-based telephone. People are less likely to phone someone in Skype because they believe they may be intruding, whereas an avatar walking around in Second Life would appear to be free to talk.

2.10 Quaternions

Integration of spatial audio into Second Life will require the positions and orientations of the avatars in the region for azimuth and range calculations. Second Life uses a quaternion number system to represent rotations in three-dimensional space. These quaternions will need to be transformed into an Euler rotation sequence before azimuth calculations can be done.

Quaternions are a four-dimensional number system that can represent three-dimensional rotations without the singularities that exist in Euler rotation sequences [100]. A quaternion, $q = q_0 + \bar{\mathbf{q}} = q_0 + \hat{\mathbf{i}}q_1 + \hat{\mathbf{j}}q_2 + \hat{\mathbf{k}}q_3$, is the sum of a scalar part and a vector part. A quaternion can be transformed into the Euler aerospace sequence, shown in Figure 2.17,

$$\begin{aligned} \psi &= \arctan\left(\frac{m_{12}}{m_{11}}\right) \\ \theta &= \arcsin\left(-m_{13}\right) \\ \phi &= \arctan\left(\frac{m_{23}}{m_{33}}\right) \end{aligned}$$

where

$$\begin{aligned} m_{11} &= 2q_0^2 + 2q_1^2 - 1 \\ m_{12} &= 2q_1q_2 + 2q_0q_3 \\ m_{13} &= 2q_1q_3 - 2q_0q_2 \\ m_{23} &= 2q_2q_3 + 2q_0q_1 \\ m_{33} &= 2q_0^2 + 2q_3^2 - 1. \end{aligned}$$

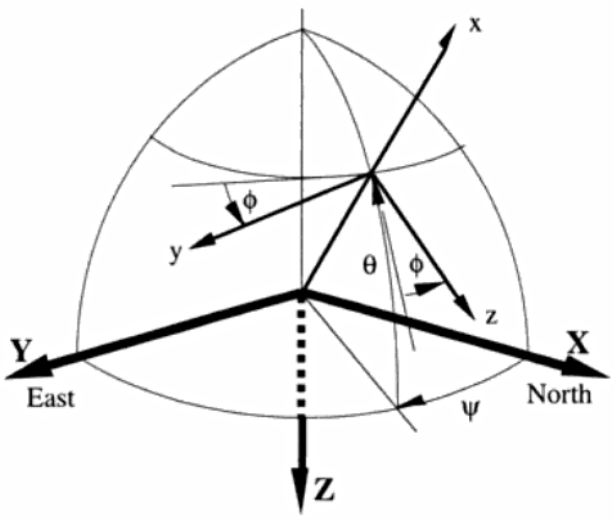


Figure 2.17: *Euler aerospace sequence (reproduced from [100]).*

The Euler aerospace rotation sequence is a *zyx* Euler sequence commonly used in aircraft and aerospace applications.

- First, a rotation through the angle ψ about the z-axis defines the aircraft heading.
- Followed by a rotation about the new y-axis through an angle θ defining the aircraft elevation.
- Finally, the aircraft bank angle ϕ, is a rotation about the newest x-axis.

The x-axis points forwards, the y-axis to the right and the z-axis downwards.

2.11 Summary

A study of existing literature in the fields of psychoacoustics and audiology relating to spatial hearing was done. The literature states that our ability to hear in noisy environments is largely due our stereophonic hearing system using spatial cues arising from the different paths that sound travels to each ear to locate and separate sound sources. This phenomenon is commonly called the "cocktail party" effect is the subject of much research. The literature suggests that simulating spatial audio will make multiple participant conversations easier to follow. A number of different techniques for generating spatial audio were discussed, with the chosen technique, HRTF spatialisation being discussed in more detail. Some existing implementations of spatial audio in entertainment and communication were discussed. The VoIP communications architecture was chosen as the most suitable for the project. Fundamental acoustics that are critical to the successful implementation of a spatial audio telephony system, such as acoustic models and HRTF interpolation, are discussed. An overview of virtual worlds in general, specifically focussing on Second Life, is given.

Chapter 3

Applications of Spatial Audio

Spatial audio has many different applications and a few will be discussed in this chapter. Section 3.1 on page 30 will give the high level design for the spatial audio telephony application that will be further developed in Chapter 6 on page 71.

A few of the applications of spatial audio, along with their supporting technologies, are given in Figure 3.1. Audio spatialisation is for taking audio that exists in the virtual domain into the physical realm, and audio localisation is for bringing audio that exists in the physical realm into the virtual domain. This can be used to create conversational environments that are based in both virtual and physical worlds. The same acoustic source separation and localisation used to render an environment into the virtual domain can be used to create microphones that track a speaker's voice and do not require close proximity. Spatial audio can be used for security and home automation by taking the point of origin of an acoustic event into account as well as the content of said event. Automated visual monitoring of a room is a complex task, but responding to auditory events can be done if the acoustics in the environment have been accurately modelled. Two of these applications will now be discussed in more detail in the remainder of this chapter.

3.1 Speech-based Chat Room

Telephony applications generally just mix the audio streams together in conversations with more than two users. This application aims that make such a scenario more immersive for the users.

A virtual environment exists and the users are free to move around in said environment, starting up and ending conversations much in the same manner as one would in a café or the cocktail party alluded to by Cherry [58]. The conversations are initiated in a dynamic manner, users do not dial each other as they would with a traditional phone of VoIP system, they simply move closer to the users they wish to engage. When users are within auditory range they will be able to communicate with each, forming dynamic conversations as are shown in Figure 3.2. Sound from users outside hearing range would either be muted entirely or brought down to a level of ambient background chatter. Keeping the sounds from other

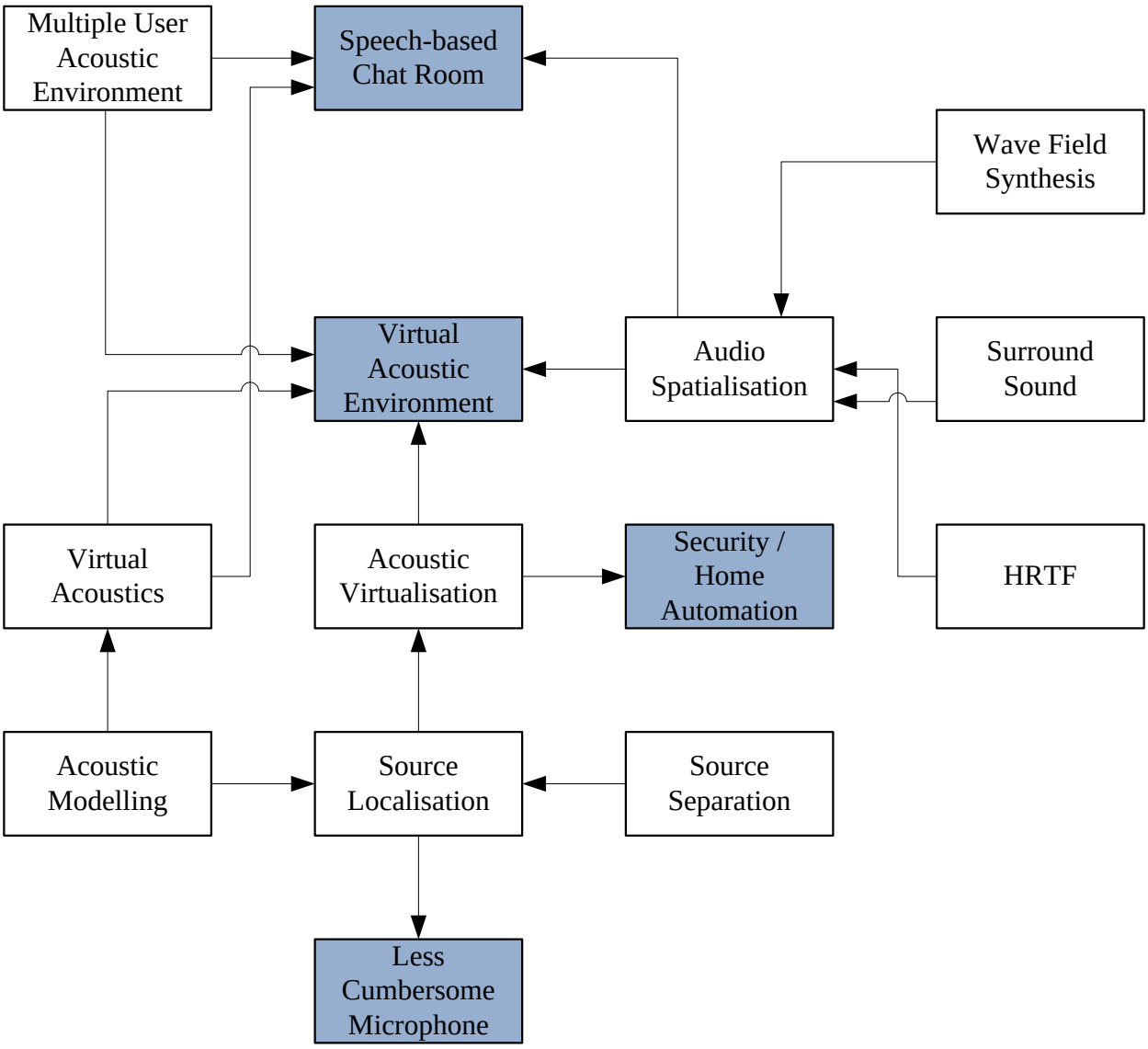


Figure 3.1: *Applications (shaded in blue) and technologies surrounding spatial audio.*

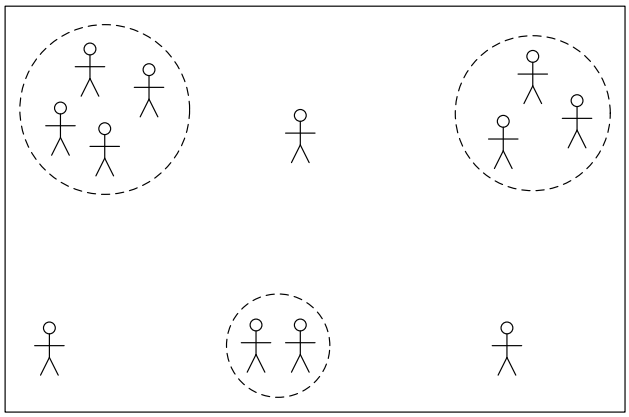


Figure 3.2: *Voice chat room with dynamic conversations.*

conversations as unintelligible background noise would make the experience more akin to that found in a café.

It could also be seen as a audio implementation of a text-based chat environment like Internet Relay Chat (IRC). A phone call, VoIP call or even a conference call using either of the aforementioned technologies does not have the same user participation model as IRC. IRC does not require a user to devote their full attention to the conversation that is underway in the way that a telephone call does, users can follow in the background and add to the conversation as they wish.

The system should be implemented entirely in software, making hardware additional to what is normally needed for VoIP calls mandatory will stunt mass adoption.

3.1.1 Spatial Audio Models

For the application described in Section 3.1 on page 30 some form of positional audio model will need to be implemented. A number of approaches exist for implementing spatial audio and we will give a brief overview of three such approaches. The choice of model will depend on the purpose of the application and the resources available to implement said application.

Binary

The simplest positional audio model is a binary one, the sound is toggled on and off depending on the distance between source and listener. Figure 3.3 shows such a model. The distance between source and listener is designated as d. The listener is connected to the audio of the source if $d \leq a$ and disconnected if $d \geq b$, the hysteresis ensures their is no instability encountered when $d = a$. Users can form dynamic group conversations by moving near each other. The model does not take any acoustic propagation effects into account.

This model provides a means for users to engage in dynamic conversations, but will not give any auditory separation between voices and does not provide for an immersive experience.

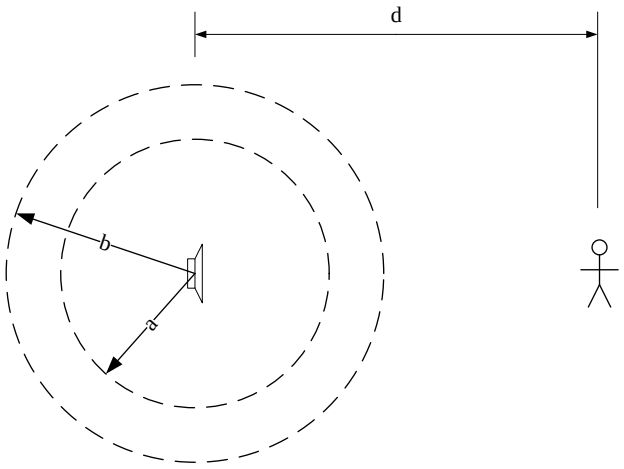


Figure 3.3: *Binary positional audio model.*

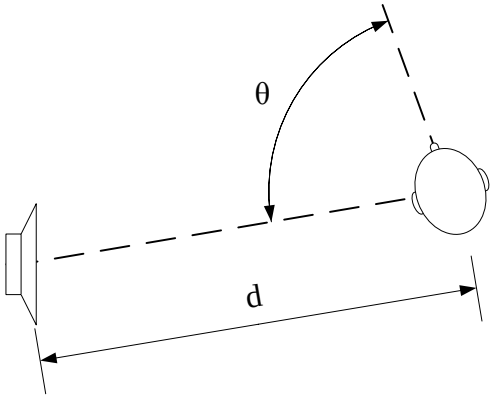


Figure 3.4: *Directional audio model.*

Directional

This model considers both the distance between the sources and the angle of the source relative to the listener as shown in Figure 3.4. The intensity of the sound is related to the distance d, increasing as the listener approaches the source and decreasing as the listener retreats. The audio will appear to originate from the azimuth θ using either HRTF spatialisation or a simpler stereo intensity model.

Full

More comprehensive acoustic modelling is necessary for applications where realism is key, such as games or training simulations.

Headphone HRTFs will be necessary due to the demand for accurate spatialisation. To model room acoustics, a reverberation FIR filter should be implemented. Simulating reverberation impulse responses is an intensive task. This can be done offline and a database with an impulse response $g_{S,M}$ for each pair of source and listener positions, (S, M) pre-generated.

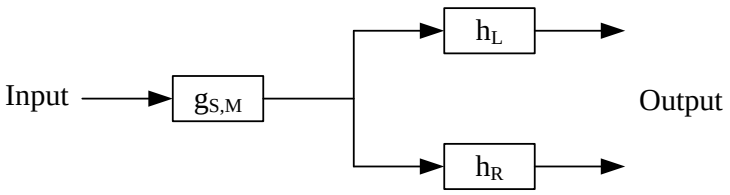


Figure 3.5: *Processing audio with reverberation and spatial audio.*

Depending on the size of the room and the sampling interval such a database could require a significant amount of storage space, but it would conserve processing power. The plenacoustic sampling theory discussed in Section 2.8.2 on page 25 will give the necessary spatial sampling interval for accurate reconstruction.

A block diagram summarising the audio processing is shown in Figure 3.5. Each monaural audio stream will first be processed by the room impulse response $g_{S,M}$ giving a monaural signal that is processed by a pair of HRTFs giving the stereo output that contains both the room response and binaural localisation cues.

This approach will be computationally expensive, possibly forcing a hardware-based solution, but gives the most realistic audio out of the three approaches considered.

3.2 Augmented Reality Auditory Environment

The spatial audio application discussed so far has focused on giving monaural sounds, that have no inherent position in the physical realm, a perceived position in the virtual realm. HRTFs, distance and reverberation modelling can make a sound appear to come from a certain direction. An inverse problem exists, namely taking a real auditory scene and breaking it up into elementary sound sources and their positions as well as characterising the room impulse response. Auditory scene analysis such as this would allow for a multitude of new applications, with one example being given in the remainder of this section. The idea proposed in this section is a very early concept for a system that aims to "virtualise" an acoustic environment and but will be a costly and complex system to implement, well beyond the scope of this project.

A number of microphones and loudspeakers are placed in an environment, ideally both will be integrated into a single module. The microphone network, using acoustic source separation and tracking [52, 81, 109, 117, 132, 130], will segregate the sounds in the room into streams and positional information. The room acoustics will also need to be measured. This would then allow remote users to take a virtual walk in the room and hear what they would have heard if they were actually there. In a similar fashion the loudspeaker network will be able to pan a virtual source around the room. The system allows for true auditory telepresence. Installing the system in two geographically separated rooms, as shown in Figure 3.6, will allow people to have augmented reality conversations with each other over

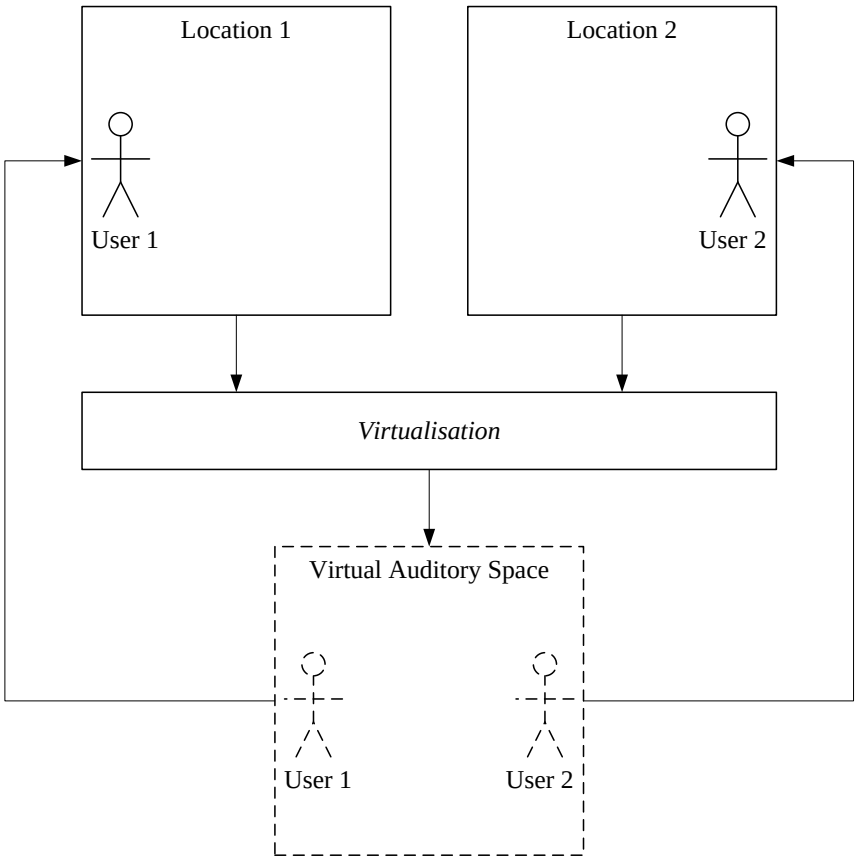


Figure 3.6: *Virtual auditory space.*

the Internet and further transcend their physical boundaries.

3.3 Summary

This chapter gave an overview of some possible applications of spatial audio. The conceptual design of a conversational audio environment implementing spatial audio was given in Section 3.1 on page 30. The application could be seen as the auditory equivalent to IRC or as a sort of "virtual coffee shop", with users entering conversations in a dynamic manner. The practical design and implementation of this application will follow in Chapter 6 on page 71. A augmented reality auditory environment was conceptualised in Section 3.2 on page 34. This application would allow two geographically separated parties to, from an auditory perspective, experience each other being in the same room.

Chapter 4

Theoretical Development

This chapter details the theoretical work that needs to be done before starting on any practical implementation of the system and experiment design that follows in Chapters 5 and 6 on pages 50 and 71 respectively.

4.1 Acoustic Model

Any application that tried to implement audio more advanced than a basic mixing to monaural will require an acoustic model, the complexity of which will be dependant on the chosen application and available resources.

4.1.1 Reverberation

The image method model for simulating reverberation, discussed in Section 2.8.1 on page 22, does not provide any indication of what the length of the resulting impulse response will be. The model will be implemented in this section to determine this length and whether reverberation is worthwhile implementing in a real-time audio platform.

The image method model was implemented as a discrete-time FIR filter. The infinite summation in Equation 2.5 across r is implemented as three finite summations across i, j and k. The discrete-time impulse response is

$$h[n] = \sum_{p=0}^{7} \sum_{i=-W}^{W} \sum_{j=-W}^{W} \sum_{k=-W}^{W} \beta_{x1}^{|i-d|} \beta_{x2}^{|i|} \beta_{y1}^{|j-e|} \beta_{y2}^{|j|} \beta_{z1}^{|k-f|} \beta_{z2}^{|k|} \times \frac{\delta\left[t - f_s \left|R_P + R_r\right| / c\right]}{4\pi \left|R_P + R_r\right|}, \quad (4.1)$$

where f_s is the sampling rate.

The impulse response is simulated for a source at (1 m, 8 m, 1.8 m), a listener at (9 m, 3 m, 1.8 m) in a room of dimensions 10 m x 10 m x 2.5 m with a reflection coefficient of 0.9 for the walls and 0.7 for the floor and ceiling. The sampling rate, f_s, is 44.1 kHz and the summations were limited to $W = 10$. The impulse response is shown in Figure 4.1. The impulse response has a long tail of 881.86 milliseconds, translating to a FIR filter of length 38891 samples. Such a long filter is extremely computationally expensive and not practical to implement in a real-time system. Calculating the impulse response is also a

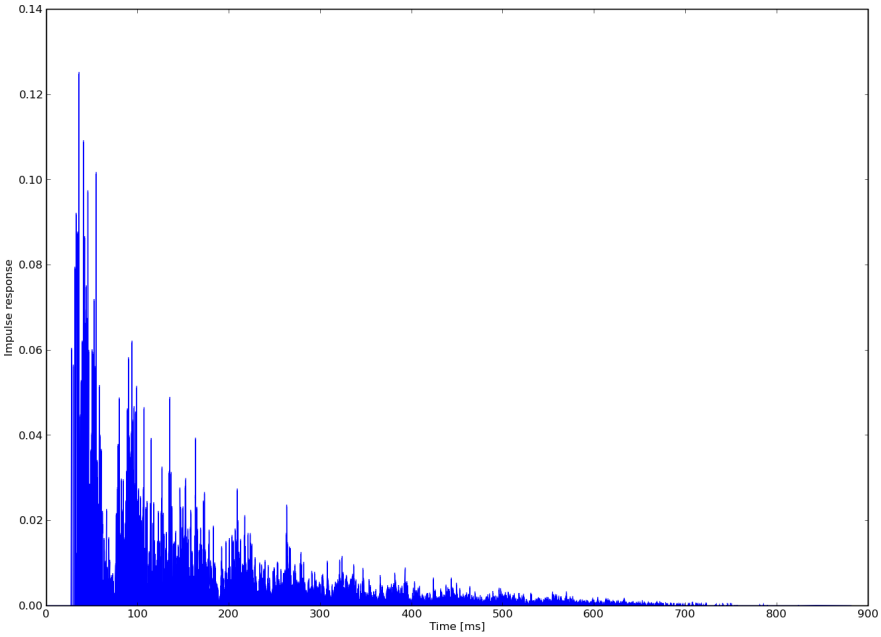


Figure 4.1: *Room impulse response.*

resource intensive task, but responses for a source and listener at various positions could be determined offline and stored for later use. For a system where realistic acoustics are not necessary, a simpler distance model would suffice.

4.1.2 Distance

Simulating audio attenuation over distance is important when implementing spatial audio in a virtual world application, where a large number of possible audio sources could be in the same region as the user. If all sources have equal loudness, regardless of distance, the resulting cacophony will overwhelm the user. Distance attenuation will make distant sounds inaudible.

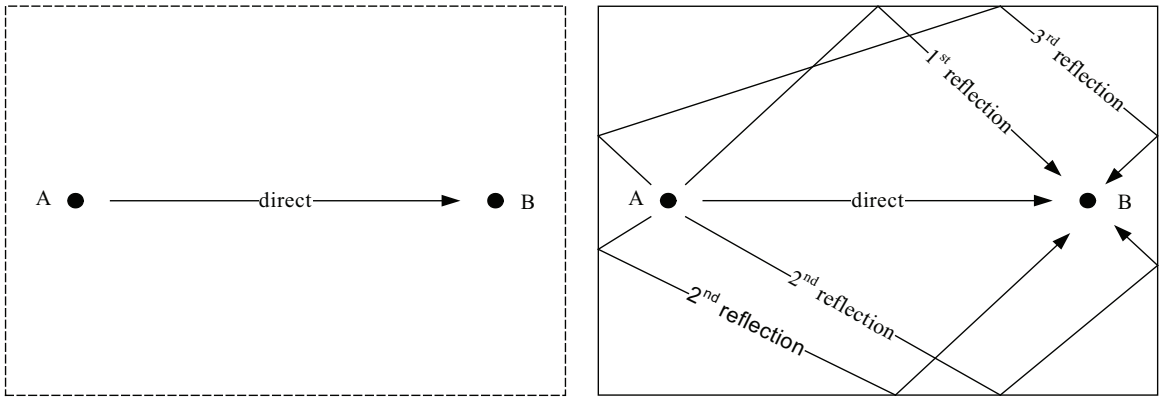
In free space, sound attenuates over distance according to the inverse square law [67], where each doubling of the distance reduces the sound intensity fourfold. The sound intensity at distance r_x from a point source is

$$I_x = \frac{I_0 r_0^2}{r_x^2}, \tag{4.2}$$

where I_0 is the reference intensity at distance r_0 from the source. The gain, g_x, is then

$$g_x = \frac{I_x}{I_0} = \left(\frac{r_0}{r_x}\right)^2. \tag{4.3}$$

Using only the inverse square to calculate the intensity of a sound at a distance r_x gives too rapid falloff of sound energy as reverberation is not taken into account. Figure 4.2 shows the difference in propagation for sound in a reverberant environment versus an anechoic



(a) Anechoic environment (b) Reverberant environment

Figure 4.2: *Sound propagation over distance.*

environment. In a reverberant environment the sound energy propagating from $A \rightarrow B$ is the sum of the direct energy and higher order reflections. The total energy will not fall off as quickly as that dictated by the inverse square law. Reverberation can be simulated [37], but the lengthy impulse responses are computationally expensive. Such a realistic acoustic model is not necessary for casual conversational environments. A model that decays less rapidally initially than the inverse square law will suffice. The model should also not have the singularity of the inverse square law at $r = 0$. An exponential model

$$g_x = 2\left(e^{\frac{2r_x}{5r_0}}\right)^{-2}, \tag{4.4}$$

that meets these requirements was found through casual experimentation. Figure 4.3 shows the gain as a function of distance for the exponential model and the inverse square law.

4.2 Acoustic Spatialisation Models

Aside from the HRTF spatialisation discussed in Section 2.4.4 on page 12, a headphone stereo panning model and a binaural model incorporating ILD and ITD will also be implemented. These models will be developed in this section.

4.2.1 Stereo Panning

Stereo panning is a technique in which a monaural signal is placed in a stereophonic sound field, setting the apparent horizontal postion of the sound by changing the output levels of the two loudspeakers. The "sine-cosine" pan law, also known as the "tangent" law, is the most common of these techniques and has a long history of use [75] and the output is given as:

$$y_L(t) = \cos(\theta)x(t) \tag{4.5}$$

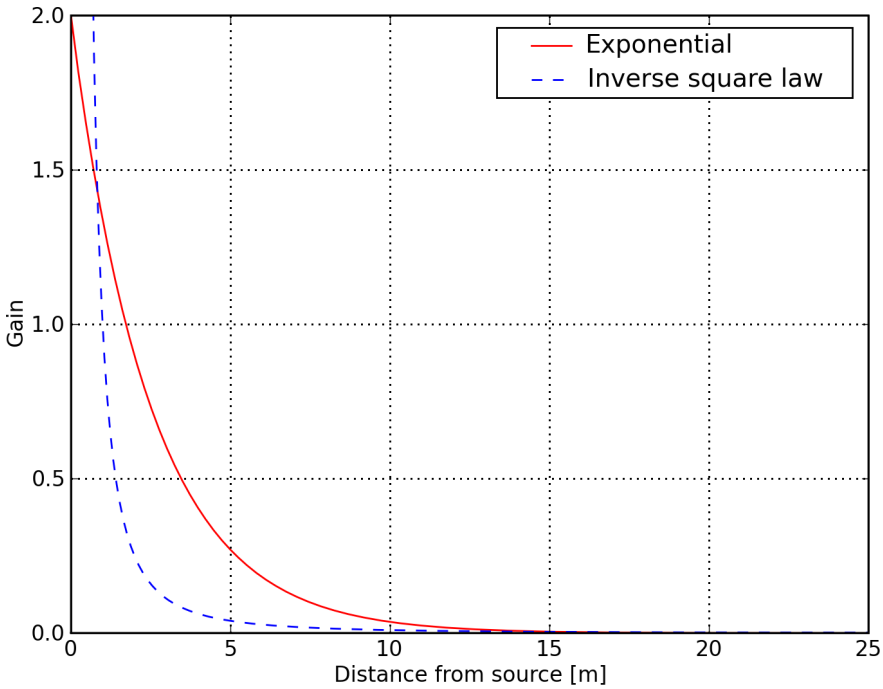


Figure 4.3: *Sound attenuation due to distance propagation.*

and

$$y_R(t) = \sin(\theta)x(t), \tag{4.6}$$

where θ varies from 0° for a source $x(t)$ panned fully to the left to 45° for a source panned to the centre and 90° for a source panned fully to the right. The sine-cosine law has the advantage of maintaining constant energy and therefore contant loudness as position is varied

$$y_L(t)^2 + y_R(t)^2 = x(t)^2. \tag{4.7}$$

The traditional sine-cosine law places the direction that the listener is facing at 45°. The traditional sine-cosine law also relies on loudspeakers that are placed 45° to the left and to the right of where the listener is facing. This does not translate properly on a system using headphones, where the "loudspeakers" are 90° to the left and to the right of where the listener is facing.

Modifying the equations to rather place the direction that the listener is facing at 0° gives

$$g_L = \cos(\theta + 45°) \tag{4.8}$$

and

$$g_R = \sin(\theta + 45°), \tag{4.9}$$

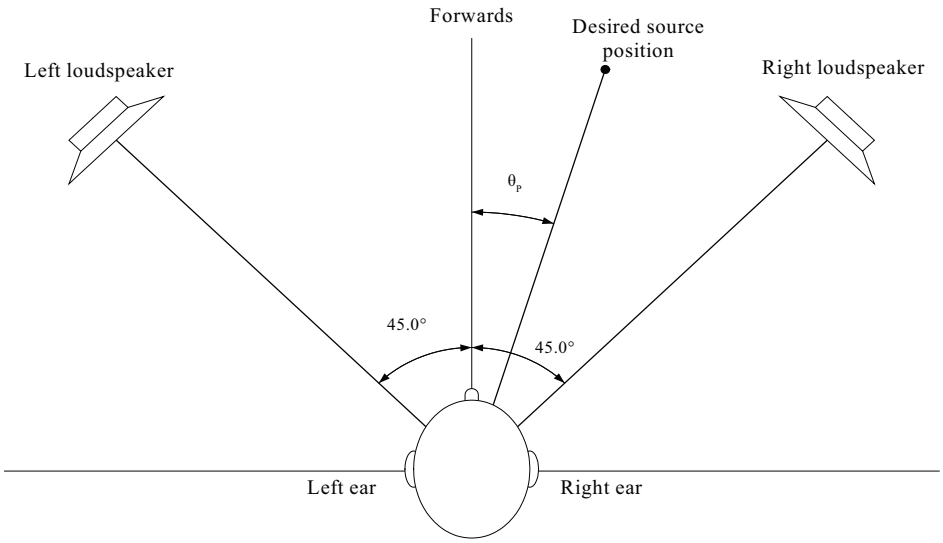


Figure 4.4: *The positions of the loudspeakers and pan angle relative to the listener.*

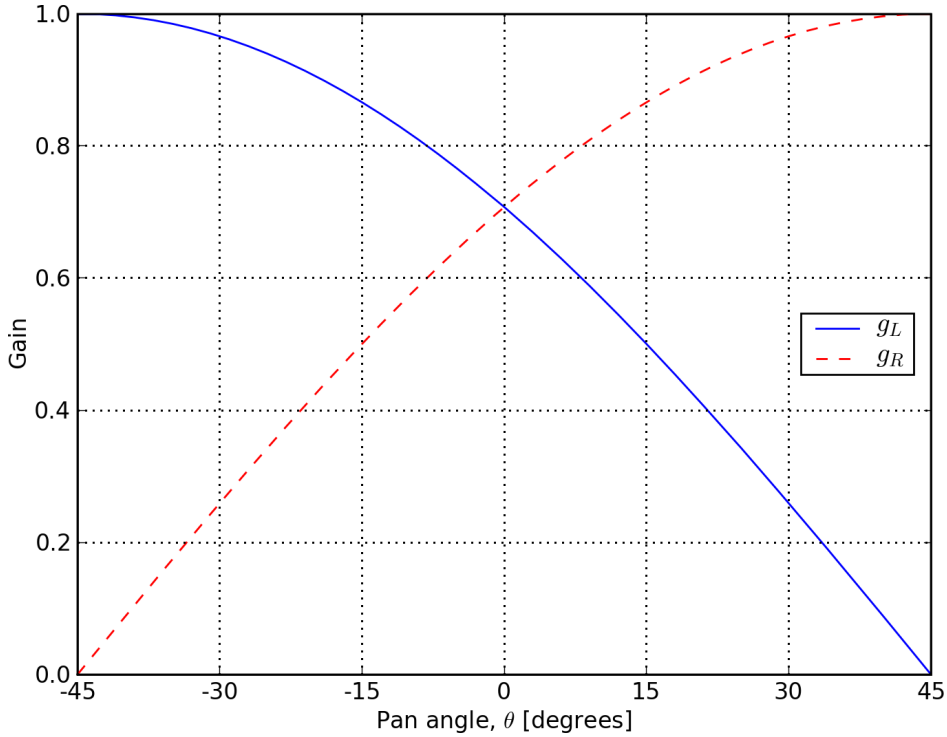


Figure 4.5: *The gain factors for a sound source at pan angle θ from the direction the listener is facing.*

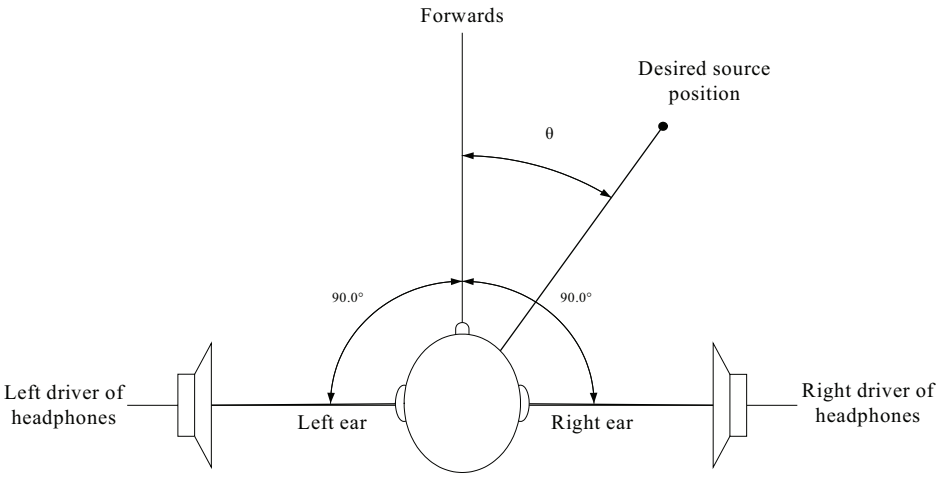


Figure 4.6: *The positions of the headphone drivers and desired source position relative to the listener.*

where θ is the pan angle and g_L and g_R the gain for the left and right channel respectively. The geometric arrangement is shown in Figure 4.4 and the gain factors are shown in Fig. 4.5.

Extending to model to work for headphones (or loudspeakers placed 90° to the left and to the right of the listener) gives

$$g_L = \cos(\theta/2 + 45°) \tag{4.10}$$

and

$$g_R = \sin(\theta/2 + 45°). \tag{4.11}$$

The geometric arrangement is shown in Figure 4.6 and the gain factors in Fig. 4.7. The audio is only played to the subject's left ear when $\theta = -90°$, to both ears when $\theta = 0°$ and only to the right ear when $\theta = 90°$. The headphones model still maintains constant energy (and therefore constant loudness) because $g_L^2 + g_R^2 = 1$.

4.2.2 Basic Binaural Model

The basic binaural model is a pure geometric model that is developed as a simplification of the HRTF model discussed in Section 2.4.4 on page 12. The model uses the different distances travelled by sound to each ear, as can be seen in Figure 2.4 on page 13, to calculate ILD and ITD functions. The model does not take into account any reflection, absorption and diffraction effects resulting from the subject's head and torso. The geometric arrangement is shown in Figure 4.8.

The distance from the source to the left ear is

$$\begin{aligned} R_L &= \sqrt{R_0^2 + \left(\frac{d}{2}\right)^2 - 2(R_0)\left(\frac{d}{2}\right)\cos(\psi)} \\ &= \sqrt{R_0^2 + \frac{d^2}{4} - R_0 d\cos(90° + \theta)} \end{aligned} \tag{4.12}$$

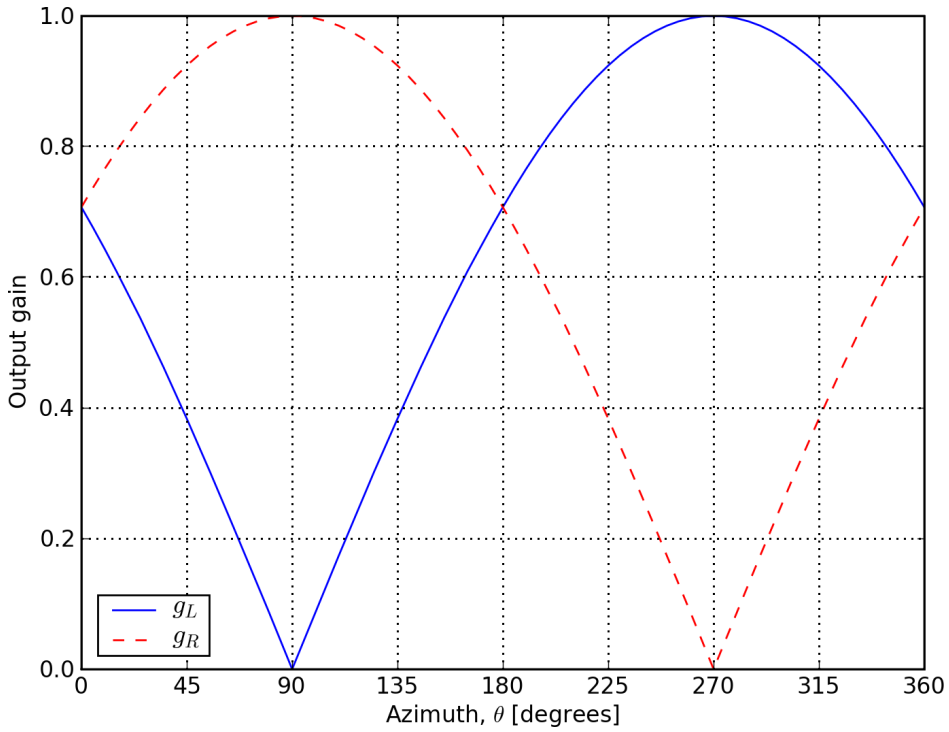


Figure 4.7: *The gain factors for a sound source panned at angle θ from the direction the listener is facing.*

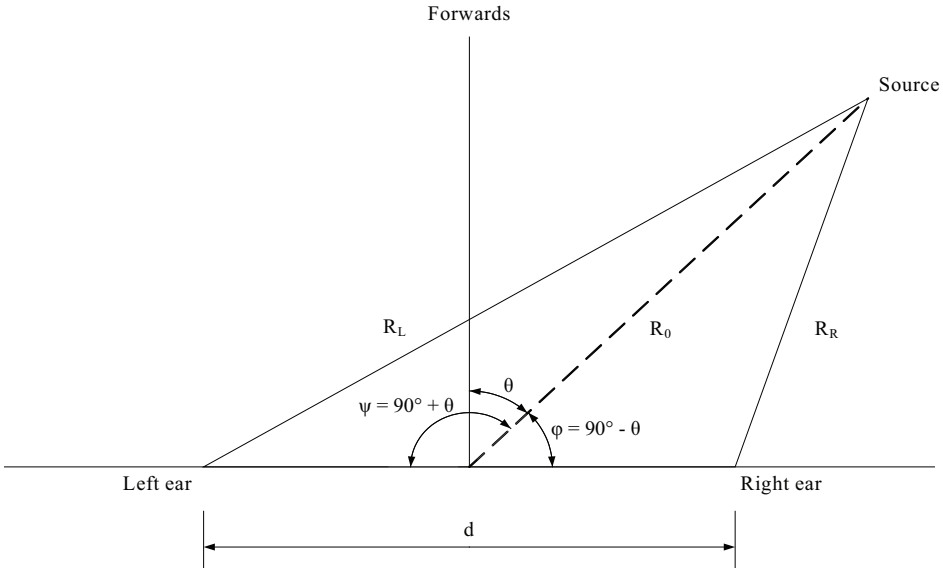


Figure 4.8: *Geometry of basic binaural model, looking from the top.*

and the distance to the right ear is

$$\begin{aligned} R_R &= \sqrt{R_0^2 + \left(\frac{d}{2}\right)^2 - 2(R_0)\left(\frac{d}{2}\right)\cos(\phi)} \\ &= \sqrt{R_0^2 + \frac{d^2}{4} - R_0 d\cos(90^\circ - \theta)} \end{aligned} \tag{4.13}$$

where $R_0 = 1.95$ m is the distance from the source to the centre of the listener's head (chosen to be the same distance that the HRTFs from the Listen database are measured at), R_L and R_R are the distances from the source to the listener's left and right ears respectively and the inter-aural distance (spacing between the two ears) is taken as $d = 18$ cm [53]. The azimuth of the sound source is θ, measured clockwise with $\theta = 0^\circ$ in the direction that the listener is facing.

Attenuation Due to Distance

Sound attenuates over distance according to the inverse square law [67] and the gain is given in equation 4.3. Using the distances from equations 4.12 and 4.13 gives the gain for the left channel as

$$\begin{aligned} g_L &= \left(\frac{R_0}{R_L}\right)^2 \\ &= \frac{R_0^2}{R_0^2 + \frac{d^2}{4} - R_0 d\cos(90^\circ + \theta)} \end{aligned} \tag{4.14}$$

and the gain for the right channel as

$$\begin{aligned} g_R &= \left(\frac{R_0}{R_R}\right)^2 \\ &= \frac{R_0^2}{R_0^2 + \frac{d^2}{4} - R_0 d\cos(90^\circ - \theta)} \end{aligned}, \tag{4.15}$$

which are the contributions of the ILD to the output. The gain can have a value larger than unity and needs to be normalised to ensure that clipping of output audio files does not occur. Looking at the function for the left ear, the maximum occurs when the source is the closest to the left ear, when $\theta = -90^\circ$. The maximum gain is

$$g_{max} = \frac{R_0^2}{R_0^2 + \frac{d^2}{4} - R_0 d}. \tag{4.16}$$

Time Delay Due to Distance

The time delay resulting from sound propagation over a distance is $\tau = r/c$, where r is the distance travelled and $c = 343$ m/s the speed of sound. Therefore the time delays for the left and right channels are $\tau_L = R_L/c$ and $\tau_R = R_R/c$ respectively, which are the contributions of the ITD to the output.

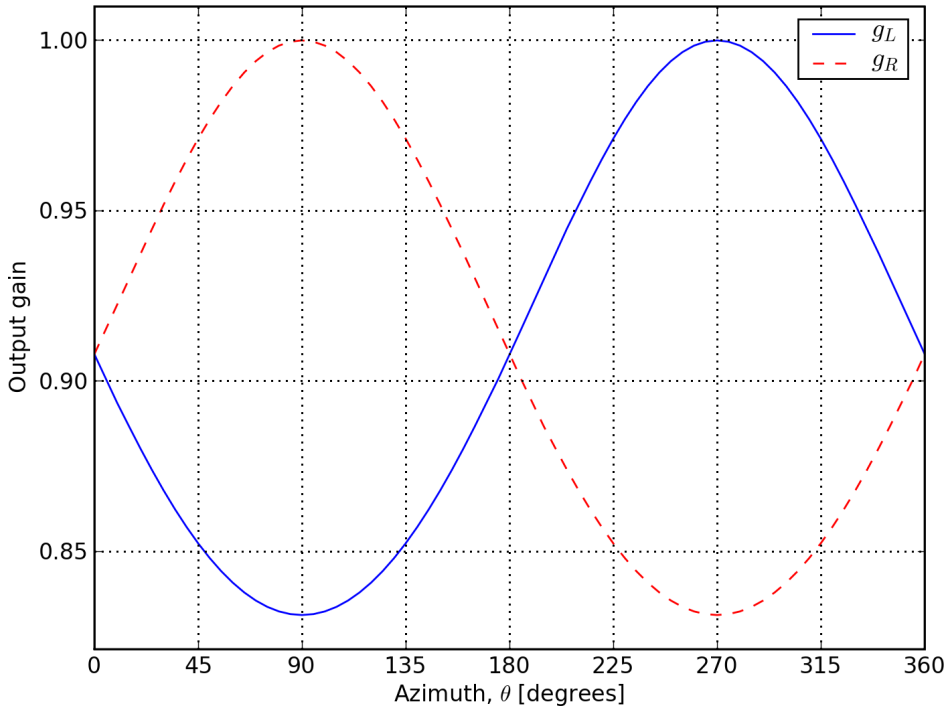


Figure 4.9: *ILD for a sound source at angle θ from the direction the listener is facing.*

Audio Model

Combining the effects of the normalised ILD and the ITD , when the input sound source is a monophonic signal x_t, the sound signal heard by the left ear is

$$\begin{aligned} x_L(t) &= \frac{g_L x(t-\tau_L)}{g_{max}} \\ &= \frac{\left[R_0^2 + \frac{d^2}{4} - R_0 d\cos(90^\circ+\theta)\right] x\left(t - \frac{1}{c}\sqrt{R_0^2 + \frac{d^2}{4} - R_0 d\cos(90^\circ+\theta)}\right)}{\frac{R_0^2}{R_0^2+\frac{d^2}{4}-R_0 d}} \end{aligned} \tag{4.17}$$

and the sound heard by the right ear

$$\begin{aligned} x_R(t) &= \frac{g_R x(t-\tau_R)}{g_{max}} \\ &= \frac{\left[R_0^2 + \frac{d^2}{4} - R_0 d\cos(90^\circ-\theta)\right] x\left(t - \frac{1}{c}\sqrt{R_0^2 + \frac{d^2}{4} - R_0 d\cos(90^\circ-\theta)}\right)}{\frac{R_0^2}{R_0^2+\frac{d^2}{4}-R_0 d}}. \end{aligned} \tag{4.18}$$

The ILD is shown in Figure 4.9 and the ITD in Figure 4.10.

4.2.3 Cone of Confusion

Looking at Figure 4.11, sound sources at positions A and B, which are both at the same angle θ away from the interaural axis, will produce exactly the same ILDs and ITDs because

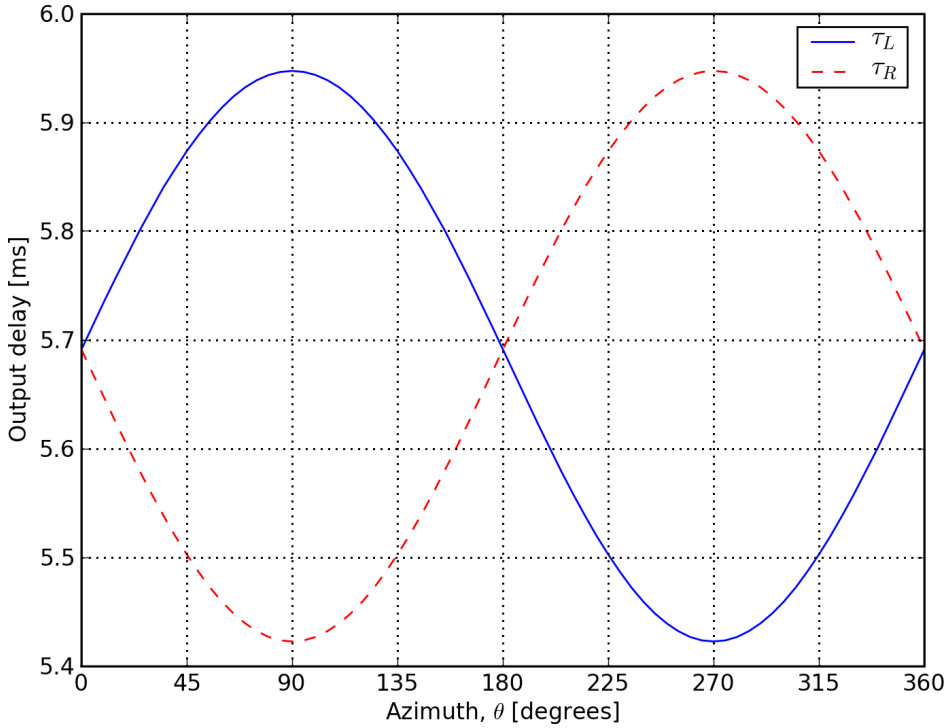


Figure 4.10: *ITD for a sound source at angle θ from the direction the listener is facing.*

only the distance of the source from each ear is taken into account. Similarly, sound sources at positions C and D will also produce exactly the same ILDs and ITDs. For this reason the stereo panning and basic binaural audio models do not provide any sense of front-back discrimination and lead to front-back confusion. This means that a sound source placed behind a listener will have the same apparent position as a source placed in front of the listener, if they are at the same angle front the interaural axis.

Localisation models that only take into account the azimuth of a sound source (limiting the position of the source to a two-dimensional plane parallel to the ground) give rise to front-back confusion. When the elevation of the source is also taken into account the exact same localisation cues will result from sources position at any point on the conical surface AB, which extends out from the listener's ear, as shown in Figure 4.12. In literature, this phenomenon is know as the "cone of confusion" [47].

4.3 HRTF Interpolation

Interpolation of the HRTF database is necessary to simulate directions not catered for in the original set and to provide a smooth aural transition between positions when simulating a moving source.

If a source is moving from one position to another it should do so as a series of small, discrete steps and not as one continuous movement. Movement in one large step as shown in Figure 4.13(a) would result in a jarring auditory transition due to the abrupt change of

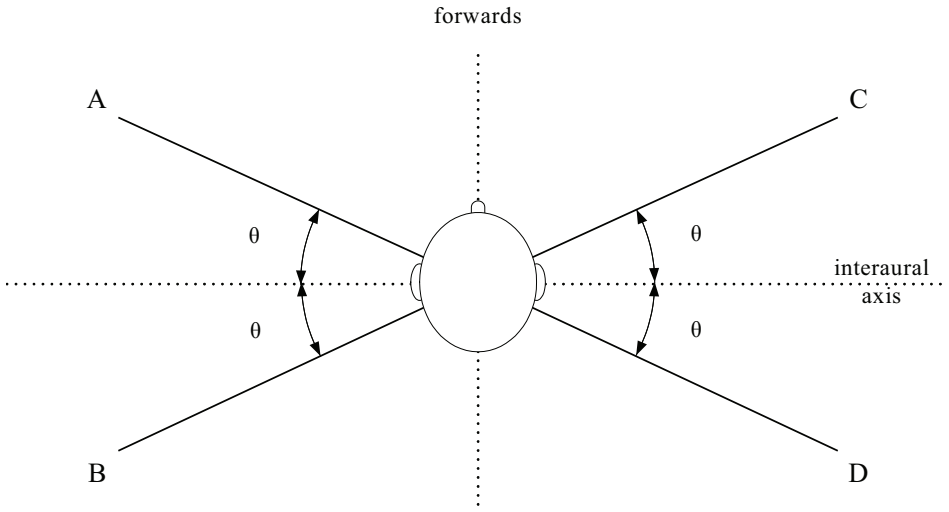


Figure 4.11: *This figure shows two pairs of sound sources, one pair on each side of the listener's head, that will produce exactly the same ILDs and ITDs for a listener positioned at the origin.*

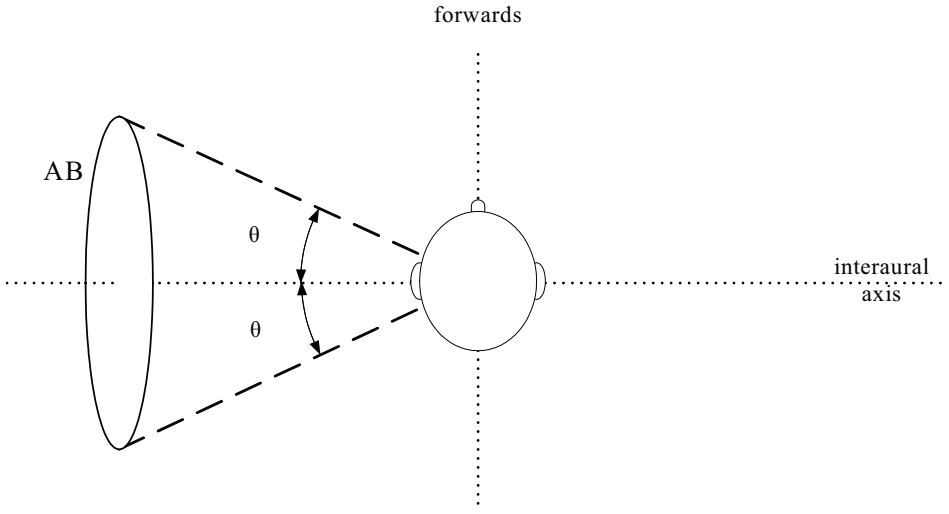
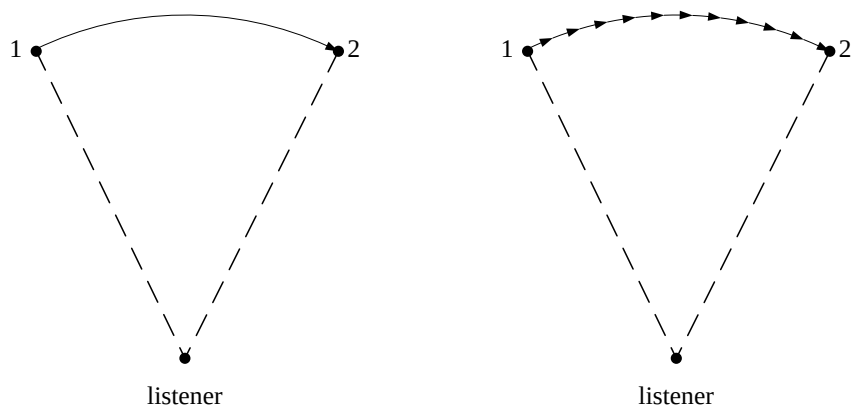


Figure 4.12: *All points on the surface AB will produce exactly the same ILDs and ITDs for a listener positioned at the origin.*



(a) In one movement (b) In small steps

Figure 4.13: *Source moving from position 1 to position 2.*

HRTF while changing the position in a number small steps as shown in Figure 4.13(b) would allow the HRTF to change in a much smoother manner. Additionally if the movement is made in one step, the sound will appear to "jump" – something that would sound unnatural to a listener.

Furthermore, in order to give a smooth auditory transition between azimuth positions, a block convolution model cannot be used for spatialisation. The taps of the filters need to be changed gradually as the position changes so that the output is not adversely affected by the different initial propagation delay due to a different position.

4.3.1 Linear Interpolation

From the literature discussed Section 2.8.3 on page 27 it is clear that interpolation will need to be done in two parts, interpolating the magnitude and time response separately. The magnitude response needs to be time-aligned before interpolation. The process for the right ear's HRIRs is exactly the same as for the left ear's HRIRs, which will now be detailed. Let h_{ϕ_1} and h_{ϕ_3} be the known HRIRs for azimuth ϕ_1 and ϕ_3 respectively and h_{ϕ_2} the unknown HRIR for an azimuth ϕ_2, which is halfway between azimuths ϕ_1 and ϕ_3. The cross-correlation [115],

$$R_{h_{\phi_1},h_{\phi_3}}[m] = \sum_{k=-\infty}^{\infty} h_{\phi_1}[k]h_{\phi_3}[k+m], \tag{4.19}$$

is used to calculate the Time of Arrival (TOA) difference between h_{ϕ_1} and h_{ϕ_3}. The TOA difference is the value of m for which $R_{h_{\phi_1},h_{\phi_3}}[m]$ is a maximum. The magnitude response of the HRIR is calculated by taking the average between the time-aligned HRIRs at ϕ_1 and ϕ_3,

$$h_{\phi_{2,mag}}[k] = \frac{h_{\phi_1}[k-m] + h_{\phi 3}[k]}{2} \text{ if } m \geq 0 \tag{4.20}$$

or

$$h_{\phi_{2,mag}}[k] = \frac{h_{\phi_1}[k] + h_{\phi 3}[k-m]}{2} \text{ if } m < 0. \tag{4.21}$$

Interpolating the ITD by delaying the magnitude response by half the TOA difference gives,

$$h_{\phi_2}[k] = h_{\phi_{2,mag}}\left[k - \frac{|m|}{2}\right]. \tag{4.22}$$

4.3.2 Interpolation and Azimuth Subdivision of HRTF Set

The interpolation algorithm can be used to generate an HRIR halfway between each pair of HRIRs again and again until the required azimuth subdivision is achieved. The nature of the spatial audio application and motion of the listener and sources in said application will determine the degree to which this azimuth subdivision is done. To increase the accuracy of the TOA determination, the HRIRs are resampled to $10f_s$, where f_s is the original sampling rate. The set of azimuth values available after the interpolation process will most likely not be exactly the same as the required azimuth set, $\mathbf{Az}$. The interpolation and azimuth subdivision process continues until there exists a subset, $\widehat{\mathbf{Az}}$, from the set of subdivided azimuths that gives $\left|\mathbf{Az} - \widehat{\mathbf{Az}}\right| < \epsilon$, where ϵ is a chosen maximum error. After the process is complete, the HRIR set is resampled back down to f_s and saved to disk.

4.4 Audio Codec

Uncompressed audio can only be used on a high bandwidth Local Area Network (LAN). If a VoIP application is to be used over the Internet, then the audio will need to be compressed and encoded with a suitable audio codec. Speex will be considered as an example of such a codec in the remainder of this section.

4.4.1 Speex Codec

Speex is an open-source, patent-free audio compression codec designed for speech. The Speex codec has three different sampling rates, 8 kHz (narrowband), 16 kHz (wideband) and 32 kHz (ultra-wideband) and the bitrate can vary from 2.15 kbps to 44 kbps [88]. Table 4.1 shows the available bit rates for the three available bandwidths. Except for the lowest quality level of 0 (which is mostly just for comfort noise), the ultra-wideband mode encodes the 0-8 kHz part of the spectrum using the wideband mode of the same quality level and encodes the remaining 8-16 kHz part of spectrum with 1800 bps, encoding only the rough shape of the spectrum.

Table 4.1: *Available bit rates for Speex codec.*

Quality	**Narrowband (8 kHz)**	**Wideband (16 kHz)**	**Ultra-wideband (32 kHz)**
0	2,150 bps	3,950 bps	4,150 bps
1	3,950 bps	5,750 bps	7,550 bps
2	5,950 bps	7,750 bps	9,550 bps
3	8,000 bps	9,800 bps	11,600 bps
4	8,000 bps	12,800 bps	14,600 bps
5	11,000 bps	16,800 bps	18,600 bps
6	11,000 bps	20,600 bps	22,400 bps
7	15,000 bps	23,800 bps	25,600 bps
8	15,000 bps	27,800 bps	29,600 bps
9	18,200 bps	34,200 bps	36,000 bps
10	24,600 bps	42,200 bps	44,000 bps

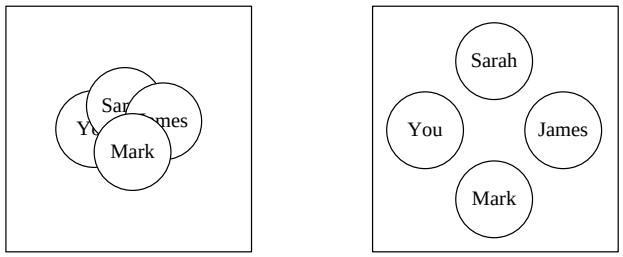
Chapter 5

System Design

This chapter gives the practical design and implementation of a spatial audio telephony system and integration of such a system into a virtual world.

One of the primary aims of our research is to demonstrate a VoIP system that could create a virtual auditory environment. Modern telephony systems use the same monaural audio model as when the telephone was first devised, which does not provide any spatial separation of speakers. This in turn makes multiple participant calls difficult to follow as the voices clump together, as shown in Figure 5.1(a). A VoIP system utilising spatial audio would have each user occupy a different position in the virtual auditory space, as shown in Figure 5.1(b). Spatial cues would allow a listener to separate the audio sources, making the conversation much easier to follow.

The spatial audio processing algorithms are detailed in Section 5.1 on page 51. Design decisions relating to the system architecture are given in Section 5.2 on page 52. The design and implementation of a telephony system that uses spatial audio is detailed in Section 5.3 on page 54. The integration of the aforementioned system into a virtual world environment, specifically Second Life, is given in Section 5.4 on page 60.



(a) Monaural voices clump together

(b) Spatial audio separates voices

Figure 5.1: *VoIP conference call.*

5.1 Spatialisation

The HRTF spatialisation is implemented by a pair of FIR filters of length N with coefficients $h_L[n]$ and $h_R[n]$. Convolving the monaural input signal $x[n]$ gives the stereo output signals

$$y_L[n] = \sum_{i=0}^{N} h_L[i]x[n-i] \tag{5.1}$$

and

$$y_R[n] = \sum_{i=0}^{N} h_R[i]x[n-i]. \tag{5.2}$$

In order to utilise the same architecture that will be developed for the HRTF spatialisation, the panning and binaural models developed in Section 4.2 on page 38 need to be implemented as FIR filters.

5.1.1 Stereo Panning

As headphone stereo panning scales the input by a pair of time-independent gain values that are a function of the azimuth which can be seen as a pair of FIR filters one sample in length. The coefficients are then

$$h_L[0] = \cos\left(\theta + 45^\circ\right) \tag{5.3}$$

and

$$h_R[0] = \sin\left(\theta + 45^\circ\right), \tag{5.4}$$

where θ is the azimuth of the source and the length of the filters $N = 1$.

5.1.2 Binaural Audio

The binaural model can be represented as a pair of FIR filters with a single impulse of magnitude the ILD and delayed by the ITD giving

$$h_L[n] = \begin{cases} \dfrac{f_s R_0^2}{R_0^2 + \frac{d^2}{4} - R_0 d\cos(90^\circ + \theta)}, & \text{if } n = \dfrac{f_s\sqrt{R_0^2 + \frac{d^2}{4} - R_0 d\cos(90^\circ + \theta)}}{c} \\ 0, & \text{otherwise} \end{cases} \tag{5.5}$$

and

$$h_R[n] = \begin{cases} \dfrac{f_s R_0^2}{R_0^2 + \frac{d^2}{4} - R_0 d\cos(90^\circ - \theta)}, & \text{if } n = \dfrac{f_s\sqrt{R_0^2 + \frac{d^2}{4} - R_0 d\cos(90^\circ - \theta)}}{c} \\ 0, & \text{otherwise} \end{cases} \tag{5.6}$$

where θ is the azimuth of the source, R_0 the range of the source, $c = 343$ m/s the speed of sound in air and $d = 18$ cm the spacing between the two ears [53]. To ensure the delays do not become too long, R_0 is limited to 1.95, as with the HRTFs, and the effect of the rest of the range is simulated through distance attenuation only. The length of the filters should be equal to the maximum possible delay in samples, therefore

$$N = f_s \frac{R_0 + d/2}{c} = 0.005947 f_s. \tag{5.7}$$

5.2 System Architecture

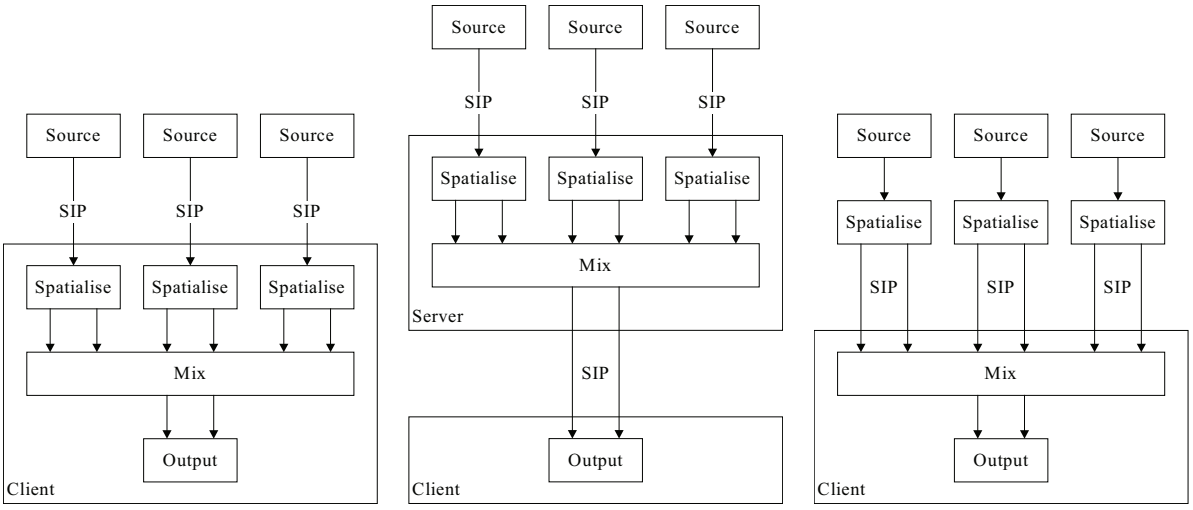
5.2.1 Client-side vs Server-side Processing

The network topology of a spatial audio VoIP system will impact computational and bandwith requirements and possible integration into existing voice-based communication systems. The network and processing power required by each method is measured in terms of the number of spatialisation operations that need to be done and a network usage coefficient. A spatialisation operation is defined as processing a single audio stream with an HRTF pair. The network usage coefficient is the sum off all audio streams incoming or outgoing at a particular node, with a stereo stream being equivalent to two monaural streams. A client-side, a server-side and a distributed approach are considered. The models are shown in Figure 5.2 from the perspective of a single client, from here on referred to as the current user. The term SIP as used includes the media stream, which is meant will be apparent by the context. The number of participants in the conversation is N and includes the current user.

The client-side approach can be dropped into a traditional VoIP system as all processing is done on the client. The number of spatialisations to be performed on each client is $N-1$. Each client has a network usage coefficient of $2(N-1)$, $N-1$ incoming and $N-1$ outgoing. The client-side approach is the least expensive to roll out as a service as it needs no additional servers or infrastructure beyond what is normally required for a VoIP service.

The server-side approach has the server perform $N(N-1)$ spatialisation operations. The server has a network usage coefficient of $3N$, N incoming and $2N$ outgoing, and each client has a network usage coefficient of 3, 2 incoming and 1 outgoing. The server-side approach does not place any additional load on the clients, but requires a powerful server for spatialisation and is therefore expensive to implement and will not scale well as a service.

The computational load of spatialisation can be minimised by a distributed approach, having each client spatialise their own outgoing audio stream and transmit it to the other clients taking part in the conversation. The clients mix together all the streams except their own. Each client has a network usage coefficient of $4(N-1)$, twice that of the client-side approach due to the use of stereo audio, but only performs a single spatialisation, that of their own outgoing stream. Each client is assigned a unique position, making this approach less flexible when compared to client-side or server-side spatialisation. This approach should work well with multicast.



(a) Client-side spatialisation (b) Server-side spatialisation (c) Distributed spatialisation

Figure 5.2: *Spatialisation topologies.*

We chose the client-side approach as it has the lowest combined bandwidth requirements and will work with existing VoIP networks. We also believe that this approach will function best in the presence of packet loss, still spatialising the degraded audio stream correctly. In the case of an approach that transmits spatial audio, packet loss could negatively affect an already spatialised stream if information essential for accurate localisation is lost. We chose not to focus too much on any specific VoIP protocol and rather work within a more general framework, using SIP as an example implementation. The position of each source could also be embedded within the SIP signalling. SIP can be substituted for any other VoIP protocol as long as signalling and media remain separate. It must be taken into account that the flexibility constraints mentioned are only applicable for the chosen topology.

5.2.2 HRTF Selection

HRTFs from the Listen database, discussed in Section 2.4.4 on page 15, were chosen to build the applications used in the research. Any other HRTF set can be used in their place, but casual experimentation suggested that they provide the best localisation accuracy with minimal pre-processing when compared to other HRTF databases.

The HRTFs from the Listen database are sampled at a temporal sampling rate of 44.1 kHz and a spatial sampling interval of 15°. The plenacoustic theory discussed in Section 2.8.2 on page 25 states that the minimum angular sampling interval necessary for interpolation is 4.95° at a temporal sampling frequency of 44.1 kHz and 6.82° at 32 kHz. It is apparent that the 15° spacing of the Listen database is not sufficient for accurate representation of angles between those measured even with interpolation. Interpolation is still necessary to provide smooth movement of simulated sources. The HRTF set was interpolated using the process outlined in Section 4.3 on page 45 with a spacing of 1° and a maximum error of $\epsilon = 0.1°$.

The sampling rate of the system was chosen to be 32 kHz to allow the use of common speech codecs, such as the Speex codec which can handle ultra-wideband 32 kHz audio [20]. Using a sampling rate not supported by an existing speech codec would require uncompressed audio to be sent over the network limiting the use of spatial audio to local networks.

5.3 Spatial Conference Call Application

It is not feasible to develop an entire VoIP system from scratch and we chose to use an existing software package as the basis for the project.

VoIP has been implemented by a number of protocols but SIP is currently the standard for multimedia communications, having been adopted by practically every public VoIP service provider [124]. Skype is a proprietary network and has a large market share, having generated 8% of international telephone traffic in 2008 [63].

5.3.1 Skype

Integrating spatial audio with Skype is not yet practical as their software is closed source because Skype does not own, or even have access to, its core peer-to-peer technology [43]. The only solution would be to use the "Skype for SIP" service developed for business users [133]. Skype for SIP is still currently in beta but would be a worthwhile option to consider for future VoIP projects.

5.3.2 PJSIP

The open source SIP stack, PJSIP [13], was used as the basis for this application. PJSIP is used for instant messaging and VoIP applications. The PJSIP project is highly portable, written in C and has a small memory footprint. These features makes it a good choice for a project that might be extended to include mobile phones and other portable media devices, environments in which resources are limited. Choosing a SIP stack that is more rigid to requirements and less efficient would sacrifice flexibility in terms of future extensions to the project. PJSIP supports stereo, an uncommon feature in VoIP because SIP can be used for applications other than just VoIP, such as audio broadcasting [27]. Version 1.0.1 of the PJSIP source code was used as the stereo implementation in earlier beta versions was not completely functional. PJSIP is of a modular nature. Figure 5.3 shows the library architecture of the PJSIP project. The two components of the PJSIP project that were modified for the purposes of the project are the PJMEDIA media stack, which does all the required media processing and PJSUA, a commandline SIP VoIP application [15]. Figure 5.4 shows a screen capture of the interface of the PJSUA commandline program.

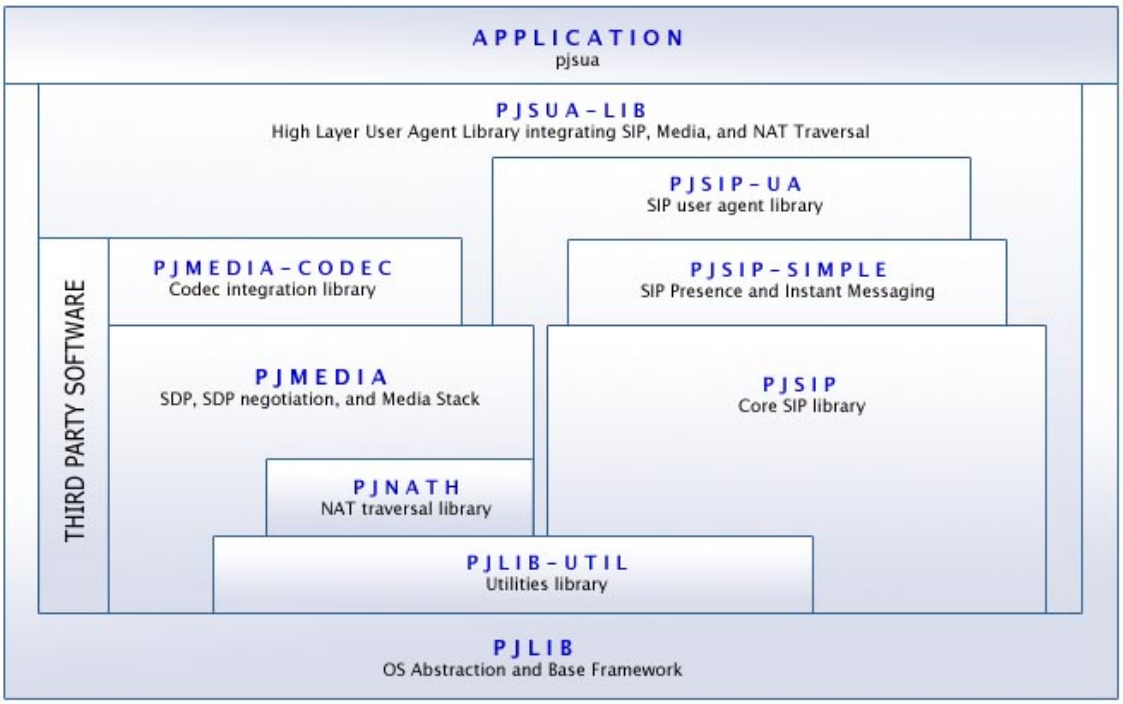


Figure 5.3: *PJSIP library architecture (reproduced from [13]).*

```
+============================================================================+
|       Call Commands:         |   Buddy, IM & Presence:  |     Account:      |
|                              |                          |                   |
|  m  Make new call            | +b  Add new buddy       .| +a  Add new accnt |
|  M  Make multiple calls      | -b  Delete buddy         | -a  Delete accnt. |
|  a  Answer call              |  i  Send IM              | !a  Modify accnt. |
|  h  Hangup call  (ha=all)    |  s  Subscribe presence   | rr  (Re-)register |
|  H  Hold call                |  u  Unsubscribe presence | ru  Unregister    |
|  v  re-inVite (release hold) |  t  ToGgle Online status |  >  Cycle next ac.|
|  U  send UPDATE              |  T  Set online status    |  <  Cycle prev ac.|
| ],[ Select next/prev call    +--------------------------+-------------------+
|  x  Xfer call                |      Media Commands:     |  Status & Config: |
|  X  Xfer with Replaces       |                          |                   |
|  #  Send RFC 2833 DTMF       | cl  List ports           |  d  Dump status   |
|  *  Send DTMF with INFO      | cc  Connect port         | dd  Dump detailed |
| dq  Dump curr. call quality  | cd  Disconnect port      | dc  Dump config   |
|  z  Call all buddies         |  V  Adjust audio Volume  |  f  Save config   |
|  S  Send arbitrary REQUEST   | Cp  Codec priorities     |  f  Save config   |
+------------------------------+--------------------------+-------------------+
|  q  QUIT       sleep MS     echo [0|1|txt]        n: detect NAT type        |
+============================================================================+
You have 0 active call
>>>
```

Figure 5.4: *Spatial VoIP application screen capture.*

5.3.3 Design and Implementation

The spatialisation is implemented entirely on the client side with no changes to the media transport protocol, making the system compatible with any existing SIP clients as well as not requiring any more bandwidth than a traditional SIP client. If the spatialisation were implemented on a central server it would be computationally expensive for the server, require at least twice the bandwidth and be incompatible with existing clients and telephony systems. Server-side spatialisation would necessitate designing a new stereo codec or interleaving stereo samples into a single monaural audio stream as PJSIP does not yet support any compressed stereo codecs [27].

A system sampling rate of 32 kHz is used as this is necessary for accurate localisation as discussed in Section 2.7 on page 19. The Speex codec was chosen as the most suitable for the system as it is the only codec supported by PJSIP that can handle 32 kHz audio, aside from the uncompressed 16 bit linear PCM codec. HRTFs from the Listen database were used after being interpolated and resampled to 32 kHz, as discussed in Section 5.2.2 on page 53.

Media ports in PJMEDIA provide a generic and extensible framework for creating media terminations [12]. A typical media port has an input audio stream, performs some processing function on this stream and delivers it as an output audio stream. An example of a media port would be the resample port built into PJMEDIA, which changes the sampling rate of the input stream. Different media ports can be connected to each other to accomplish certain functions. The modular nature of PJSIP is demonstrated by Figure 5.5, showing the media interconnection of the different components for a typical call [14]. The sound device port translates the recording and playback calls of the sound device into the *put_frame* and *get_frame* callbacks used by the media port architecture. The *put_frame* callback sends frames to downstream ports and the *get_frame* callback acquires frames from upstream ports. The conference bridge routes the audio and when multiple calls are active at the same time, mixes the audio signals. The media stream port handles the decoding and encoding of incoming and outgoing frames. The User Datagram Protocol (UDP) media transport transmits and receives RTP and RTP Control Protocol (RTCP) packets, which is driven by the flow of the network sockets and not that of the rest of the system.

The spatialisation function is implemented as a media port that is connected between the conference bridge and the stream port, processing the frames after they have been decoded by the codec port and before they are mixed together. To ensure that the sound device and conference port are initialised in stereo, a stereo port is placed before the spatialisation port. The stereo port converts the incoming monaural stream into a stereophonic stream with the same signal on both channels. No processing is done on outgoing frames as the spatialisation exists in its entirety on the client side. The spatial port processes the frames from the stereo port with a time-domain HRTFs implemented as a pair of FIR filters, giving a stereophonic output signal that gives the sound perceived direction according to an azimuth specified when instantiating the media port. The azimuth for each call is passed to the media port

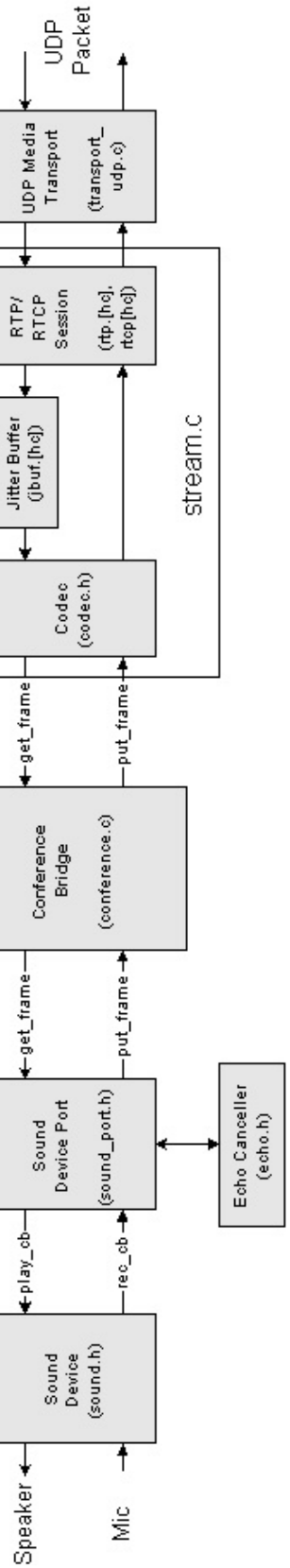


Figure 5.5: *PJSIP media flow (reproduced from [14]).*

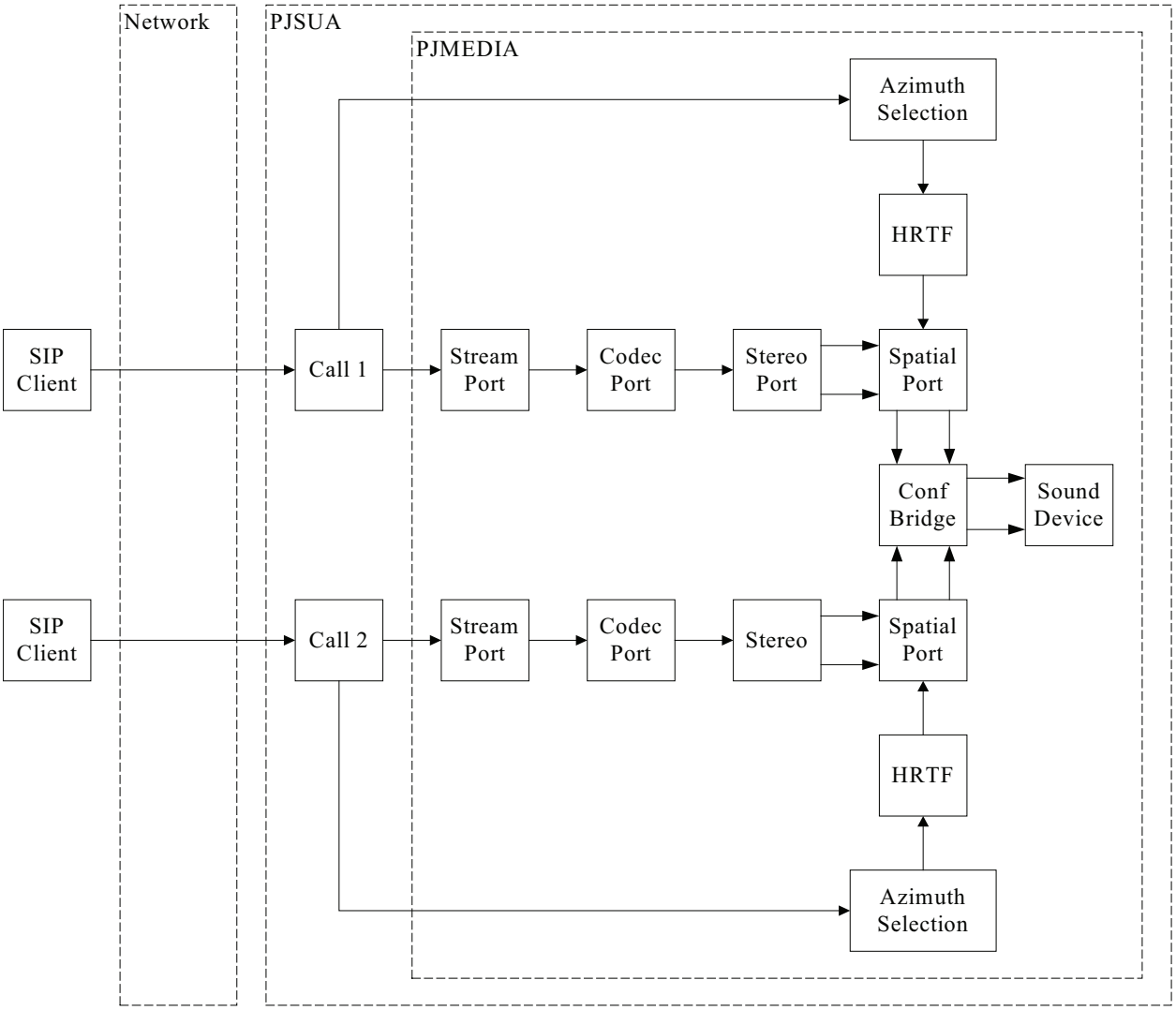


Figure 5.6: *Block diagram of spatial VoIP system showing how calls are spatialised using the media port framework of PJMEDIA.*

initialisation function by PJSUA upon creation of the call, placing each speaker at a different position. The perceived spatial position of each source is fixed during the conversation and is linked to the numerical identifier of the call. For example, call number one might be placed at an azimuth of 285°, call number two at 330° and so forth. A summary of this process for two calls is shown in Figure 5.6.

When a call is made or received, a spatial port is created, receiving an azimuth value based on the call number. Upon creation of the spatial port the HRTF coefficients for the appropriate azimuth are loaded into memory. The *put_frame* and *get_frame* callbacks of the spatial port call the spatialisation function on the audio frames they send and receive from downstream ports. The spatial port manages the initialisation and data flow of the spatialisation function. The spatial port also ensure that all memory is freed and that the spatialisation function is terminated correctly during destruction of the port upon termina-

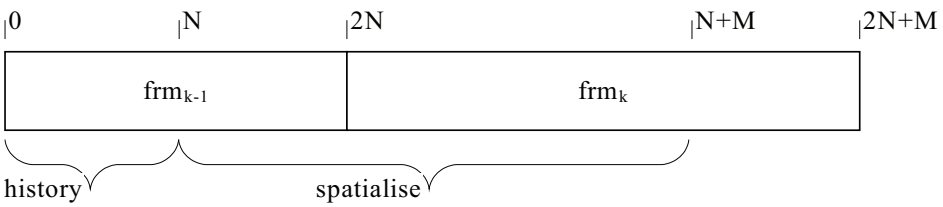


Figure 5.7: *PJMEDIA buffer for spatialisation.*

tion of the call.

PJMEDIA stores and processes all samples as 16 bit signed integers internally. To minimise quantisation errors, the spatialisation processing must be done with floating point values during accumulation and only converted right before interleaving the output samples into a single stereo stream. PJMEDIA handles audio in 20 millisecond frames. For a port set up at 32 kHz the monaural frame size is 640 samples long and the stereo frame size is 1280 samples long. The monaural frame size is designated as M and the length of the HRIR as N. The input stream is deinterleaved into two monaural frames, one for each channel. Only the frame for the left channel is used as spatialising an already stereophonic stream will negatively impact localisation. The application only provides the spatial port with a single frame on each run and the FIR filter process of the spatialisation requires N samples as history. Another N samples are needed as lookahead to provide a smooth filtering even when the filter coefficients are updated between runs. At the beginning of each run, $2N$ samples from the previous frame are copied (initialised as zero for the first run) to the beginning of the buffer and the M samples of the current frame being copied are that. The spatialisation is done on the samples in the buffer between positions N and $(N + M)$ with the first N samples of the buffer being used as history. The samples between positions $(N + M)$ and $(N + 2M)$ are to be used in the spatialisation for the next frame. This buffer arrangement is shown in Figure 5.7, with frm_{k-1} being the previous frame and frm_k the current frame. The spatialised frames are

$$y_L[n] = \sum_{i=0}^{N} h_L[i]x[(n-i)+N] \tag{5.8}$$

and

$$y_R[n] = \sum_{i=0}^{N} h_R[i]x[(n-i)+N], \tag{5.9}$$

with x being the input buffer and $n \in [0, M)$. The pair of spatialised frames are interleaved into a single monaural output frame. The nature of the buffering means that the output frame of the spatial port always lags the input frame by N samples.

The HRTF filter coefficients were truncated to $N = 325$ as buffers larger than this in PJSIP result in auditory degradation of the output. Figure 5.8 shows a stacked plot of the

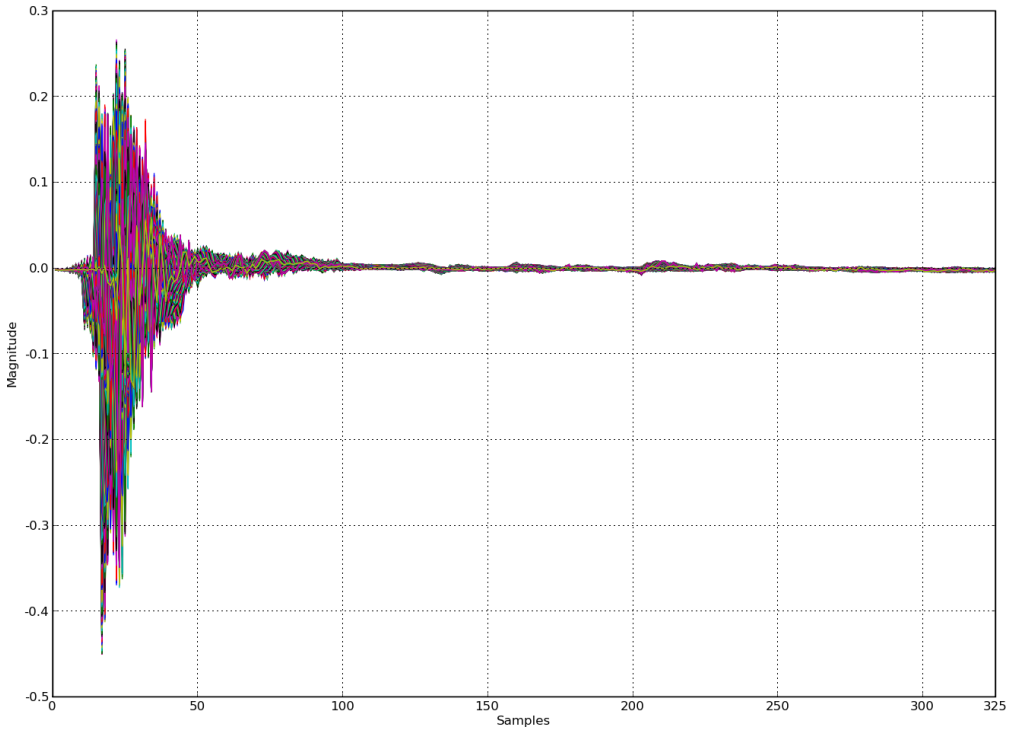


Figure 5.8: *Stacked plot of all impulse responses used in application.*

entire HRTF database used in the program, with azimuth values in 1° increments in the set $[0°; 360°)$ for both left and right channels. It can be seen that the meaningful portion of the impulse response exists before 250 samples. Limiting the impulse response to $N = 325$ will not negatively affect spatialisation.

The interface of the PJSUA client only allows the user to make calls one at a time, in order words, if the user wished to make a conference call with three other participants they would have to repeat the make call command three times. A "call all buddies" command was added that makes a call to all SIP URIs in the user's buddy list. The main purpose of a spatial VoIP application is for calls with multiple participants, and this feature makes it less effort for the user to initiate such calls.

5.4 Spatial Audio in a Virtual World Environment

The spatial VoIP application developed in Section 5.3 on page 54 requires users to actively dial anyone they wish to speak to. The conversation continues until either party chooses to hang up, the same manner in which we normally conduct telephone conversations. This is a static conversational situation. Integrating spatial audio in a virtual world, such as a Massively Multiplayer Online Game (MMOG), would provide a more free-form and dynamic

conversational situation[1]. The users would not dial each other, but merely walk into auditory range of a user they wish to speak to, more similar to how we conduct conversations in person. If the VoIP application is parallel to the traditional telephone, then the virtual world application is in essence a "virtual coffee shop", an environment in which we believe people would converse more readily in dynamic conversations. Such an application would require the azimuth and range of each call to be constantly updated based on the position and orientation of each other user relative to the current user. The avatar that the user is controlling would be the listener and every other avatar is a source.

5.4.1 Second Life and OpenSim

The Linden Labs Second Life client, also referred to as the viewer, is open source and runs on Microsoft Windows, Linux and Mac OS X. Although the official Linden Labs Second Life server and protocol is closed source, the protocol has been reverse-engineered [5], making it possible to host one's own server. OpenSimulator, often referred to as OpenSim, is a server for hosting virtual worlds [8]. OpenSim is open source and compatible with the Linden Labs Second Life client. The existence of an open source client and server, as well as their cross-platform nature, make Second Life and OpenSim an excellent platform for the development of a proof of concept application. Screen captures of Second Life are shown in Figure 5.9.

5.4.2 Spatial Interface Design

The model for interfacing Second Life is shown in Figure 5.10 for two clients. The Second Life clients communicate with the OpenSim server using the Second Life protocol. The PJSUA clients transmit audio to each other using SIP. A "world translation" module needs to be designed for each client. This module would retrieve positional information from the virtual world and send the azimuth and range of each Second Life avatar to the PJSUA client. There are a number of ways of obtaining the positional information of the avatars from the virtual world and the remainder of this section discusses some of them and their pros and cons. The positional information can either be requested from the Second Life client or the OpenSim server. A client-based approach would have an advantage over a server-based approach in terms on bandwidth usage. The OpenSim server is already serving the positions of the avatars in the region to the client, the PJSUA client requesting the positions from the server would generate redundant network traffic. Any network requests should be integrated in an external program and not directly into PJSUA as PJSUA only has a single thread for each call and any time spent waiting for requests to return would negatively impact the auditory quality of service.

---

[1]It would also be possible to create a dynamic auditory environment similar to IRC, without using a virtual world.

(a) An avatar (b) World from above

Figure 5.9: *Second Life screen captures.*

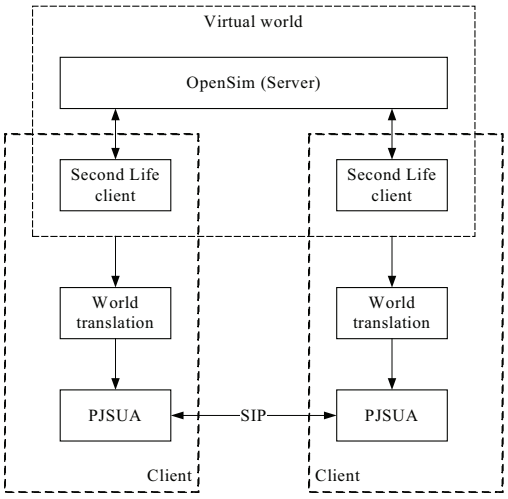


Figure 5.10: *Interfacing Second Life with PJSUA.*

Linden Scripting Language (LSL)

Linden Scripting Language (LSL) is an event driven scripting language used in Second Life to give make in-game objects accomplish certain tasks [6]. A script can be written to get the positions of all the avatars near the listener and pass them out to an external program for processing. LSL has functions that can be used to get the positions of other avatars, the position of the listener and the direction the listener is facing. The information can be sent to an external application via the Extensible Markup Language – Remote Procedure Call (XML-RPC) interface of Second Life. Second Life cannot currently initiate requests, but the *llRemoteDataReply* LSL function can be used to reply to requests from an external server. This method affords two advantages, the first is that no additional network traffic will be generated if the external XML-RPC server is run on the same machine as the client. The second being that it does not require any changes to the source code of either Second Life or OpenSim, meaning it can be used in situations where custom client or server software is prohibited. The Second Life XML-RPC interface however has two critical shortcomings making this method unsuitable for the chosen application. Calling the *llRemoteDataReply* function forces the script to sleep for three seconds, limiting the rate at which positions can be updated. The Second Life XML-RPC interface also has a high intrinsic latency, requiring approximately 300 milliseconds to reply to a request that required no complex processing. This latency was determined by packet sniffing a basic echo request using Wireshark on the local loopback network interface (localhost).

OpenSim REST Services

OpenSim uses Representational State Transfer (REST) and XML-RPC for internal and external communication [9]. REST is an architectural style, built upon Hypertext Transfer Protocol (HTTP), for distributed hypermedia systems on the World Wide Web [69]. OpenSim currently features REST requests that return information about the regions on the server. A new REST request can be written that returns the positions and rotations of all avatars in the current region. The major drawback of this method is that, the information that the client needs is requested from the OpenSim server, which generates additional network traffic. A further disadvantage of this method is that it cannot be used on the official Linden Labs servers because it requires a modified OpenSim server.

OpenSim Web Statistics Module

OpenSim has a web statistics module that provides region information [10]. This module is accessible via an Hyper Text Markup Language (HTML) page and gives the positions of all the avatars in the region. This page can be scraped by an external program and the source code for the web statistics module can be modified to return the camera rotation of all avatars as is currently done with the positions. The disadvantages of this method are the same as with using the OpenSim REST interface, it generates additional network traffic and

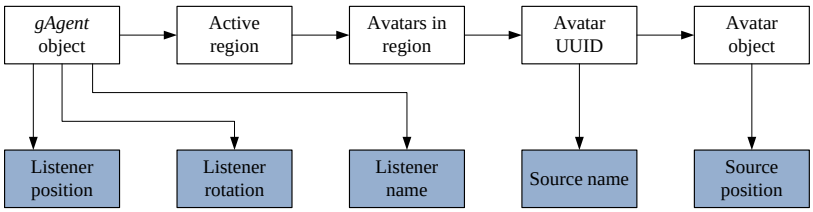


Figure 5.11: *Obtaining listener and source information from the Second Life client.*

requires a modified OpenSim server.

Custom Second Life Client

The open source version of the Linden Labs Second Life client can be modified to return the positions and rotations of all avatars in the region.

The *LLAppViewer::idle* function is called everytime the application window is not doing anything and handles general updates to the application window [17]. The end of this idle function is the most appropriate place to put code extracting positional information as it will not interfere with more time critical functions of the program.

The method by which the details of the listener and source avatar are obtained is shown in Figure 5.11, with the values of interest shaded in blue. The position and the facing direction of the user's avatar (the listener) are obtained from the *gAgent* object. The position is stored as a three-dimensional vector and the rotation as a four-dimensional quaternion. This object also has a pointer to the region that the user's avatar is currently in. From this a list of Universally Unique Identifiers (UUIDs) all other avatars (sources) in the region can be obtained. The UUID of each avatar can be used to get the name and position of that avatar.

The advantage of this method is that it does not generate any extra network traffic. Additionally, this method ensure that the perceived auditory positions will match the visual positions of the avatars on a high latency network connection where the positions of the avatars on the client might lag behind that of the actual positions on the server. The disadvantage of using this method is that it requires a modified Second Life client.

5.4.3 Spatial Interface Implementation

As all spatial processing is done client-side, the chosen approach is to modify the Second Life client to get avatar positional information and to pass it to PJSUA. This method does not require any additional network traffic, which would make the system more costly to use and using a custom client is preferable to using a custom server. However, if spatialisation were instead to be done server-side, then an OpenSim-based approach, such as using the REST interface would be best for similar reasons.

Range and Azimuth Calculation

The range and azimuth of each source, relative to the listener, is calculated within Second Life. The azimuth, elevation and range for each source can be determined from the position of the source and the heading, elevation and position of the listener. The sources are omni-directional, meaning that the direction the source is facing is not important. The elevation to the source is not used as the application is intended to be a sort of "virtual coffee shop", where the subjects are expected to remain on the same vertical plane as each other.

In Second Life, rotations are represented by quaternions, but by a different convention from that encountered in most literature. Kuipers defines the quaternion $q = \langle q_0, q_1, q_2, q_3 \rangle$, as q_0 being the scalar part and (q_1, q_2, q_3) the vector part [100]. Second Life instead defines the quaternion as (q_0, q_1, q_2) being the vector part and q_3 the scalar part [18]. Positions in Second Life regions have the positive *x-axis* pointing east and the positive *y-axis* pointing north. The quaternion to Euler angle transformation by Kuipers [100], as described in Section 2.10 on page 28, is used to calculate (ψ, θ, ϕ).

The positions of the listener and the source are (x_L, y_L, z_L) and (x_S, y_S, z_S) respectively. The range from the source to the listener is

$$\rho = \sqrt{(x_S - x_L)^2 + (y_S - y_L)^2 + (z_S - z_L)^2} \tag{5.10}$$

The angle from source to listener based on world coordinates is,

$$\gamma = \arctan\left(\frac{y_S - y_L}{x_S - x_L}\right) \tag{5.11}$$

Therefore the azimuth, shown in Figure 5.12, is

$$\alpha = \psi - \gamma = \psi - \arctan\left(\frac{y_S - y_L}{x_S - x_L}\right) \tag{5.12}$$

The elevation from the source to listener is comprised of two parts, the source elevation due to height difference

$$\zeta = \arctan\left(\frac{z_S - z_L}{\sqrt{(x_S - x_L)^2 + (y_S - y_L)^2}}\right), \tag{5.13}$$

and the direction that the listener is facing θ. Therefore the elevation, shown in Figure 5.13, is

$$\beta = \theta + \zeta = \theta + \arctan\left(\frac{z_S - z_L}{\sqrt{(x_S - x_L)^2 + (y_S - y_L)^2}}\right) \tag{5.14}$$

Call to Avatar Mapping

Being able to map a specific avatar to a specific SIP call is essential to ensuring correlation between the perceived auditory position of a source and the position of that the avatar

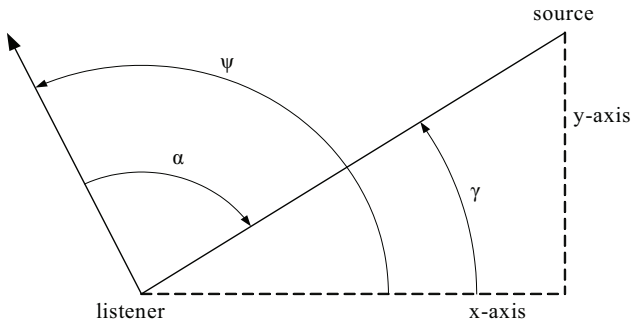


Figure 5.12: *Listener and source positions showing azimuth, α.*

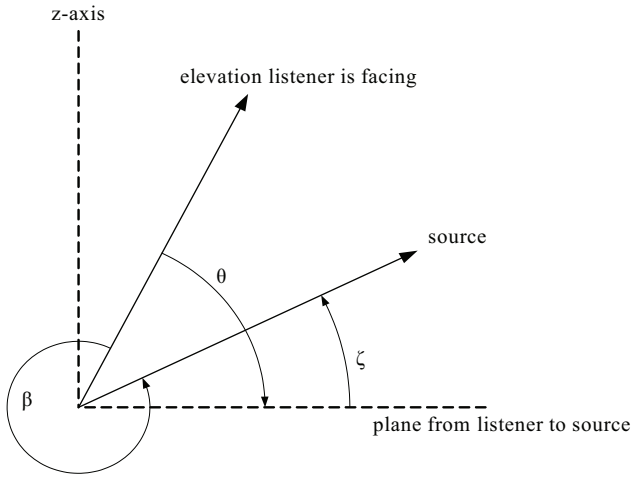


Figure 5.13: *Listener and source positions showing elevation, β.*

Listing 5.1: *Example of SIP URI to Second Life avatar name mapping.*

```
1 sip:229.42.141.251:43928 John Smith
2 sip:146.232.121.43:36581 Sarah Jones
3 sip:126.73.238.53:62912 Joe Baker
```

corresponding to that source in the virtual world. A type of "phone book" is necessary to relate Second Life avatar names to SIP URIs. Without such a system, the perceived positions of the calls would be mapped to avatars seemingly at random. The phone book is simply a file containing a phone book identifier, SIP URI and Second Life avatar name for each person that the user wishes to be able to call. Listing 5.1 shows an example of such a file, showing how the SIP URI *sip:146.232.121.43:36581* and Second Life avatar *Sarah Jones* are both mapped to the phone book identifier *2*. Second Life matches the avatar name of the current source to an entry in the phone book, retrieving the entry's phone book identifier.

Lock Implementation and Shared File

Second Life writes the phone book identifier, azimuth and range for each source in a shared file (each source on a separate line). PJSUA reads this file to obtain the azimuth and range for processing each call. A secure file locking mechanism needs to be implemented to ensure that PJSUA does not attempt to read from the file while Second Life is writing to it and vice versa as this would likely cause either program to read invalid data. The general functioning behind this locking mechanism is as such: each program tests for the existence of the other lock file, then creates its own lock file, does the check again and only then accesses the file. The lock file is deleted after this, allowing the other program access to the file. If a lock file is detected during either of the two checks, file access is skipped and the loop continues, no waiting delays which would hold up either program are implemented. The mechanism ensures that only one program can access the shared file at any given time[2].

PJSIP Modifications

To implement mobile sound sources, the application developed in Section 5.3 on page 54 needs to be modified to allow for the azimuth of a call to be changed while the call is in progress and for an attenuation proportional to the range to be applied to the input audio frames. The original application uses the call number to assign a fixed azimuth to the call upon creation of the spatial port. The modifications implemented are detailed below.

Upon creation of a new call, PJSUA matches the SIP URI of the call to an entry in the phone book and retrieves the phone book identifier of that entry. This identifier is passed

[2]This basic file locking method is suitable for the prototype application that was developed, but something more elegant would be required for a production application.

to the spatial media port created for the call and is used to get the correct azimuth and range from the shared file. When the spatial port is created the azimuth is initialised to 0° and the *pjmedia_spatial_coeffs* function is called. This is done to get the HRTF coefficient length, N as the spatialisation buffer is $2N + M$ samples long, where M us the monaural frame size. After the buffer memory has been allocated the *pjmedia_spatial_update* function is called to update the azimuth and range values by checking the shared azimuth file using the locking mechanism. The azimuth and range for the entry in the shared file with phone book identifier matching the one of the spatial port are compared to the values currently stored for the port. If they have changed, the new distance gain is calculated using the model detailed in Section 4.1.2 on page 37 as

$$g_x = 2\left(e^{0.4r_x}\right)^{-2}, \tag{5.15}$$

where r_x is the range from the source and the *pjmedia_spatial_coeffs* function is called to load the new HRTF coefficients into memory. Each audio frame is scaled by g_x after spatialisation to simulate the attenuation resulting from the propagation of sound over distance. The *pjmedia_spatial_update* function is also called just before each frame is spatialised to ensure that the most recent position is used for each source.

Figure 5.14 shows a summary of how the spatialisation process works with the file locking mechanism, where *sl.lock*, *pj.lock* and *az_file* are the Second Life lock file, the PJSUA lock file and the shared azimuth file.

The auditory transition between virtual positions is smooth because the filter taps are updated without clearing the history data in the buffer and the 1° azimuth difference between adjacent HRTFs is smaller than is detectable [125].

5.5 Summary

Spatial audio was implemented in a VoIP application, built upon the PJSIP project, allowing the user to assign each call a perceptual auditory position. The application was designed primarily with conference calls in mind. The aim is to have the perceived spatial cues create spatial separation between the different audio sources. This separation would make it easier for the listener to identify which person is currently speaking in a conference call. This would be of benefit to a participant in a contract negotiation or job interview, situations where one would answer in a different fashion when speaking to a business person than when speaking to a member of the technical team, for instance. The spatial separation between voices is also expected to increase the intelligibility of speech in a multiple source environment, ie. make it easier to listen to a specific person when more than one person is speaking at the same time. All audio processing is done on the client without any changes to the VoIP architecture, meaning that there are no additional bandwidth requirements and that the application can deployed seamlessly in any VoIP environment using SIP.

Three spatial audio models were implemented and evaluated:

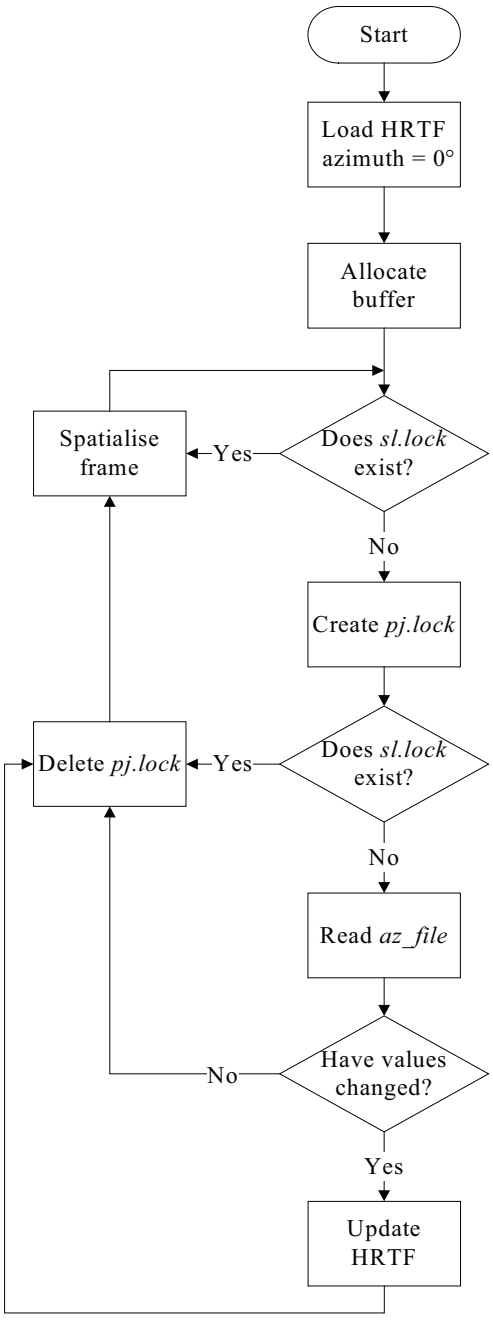


Figure 5.14: *PJSUA HRTF update process.*

- HRTF spatialisation that models the acoustic filtering resulting from different propagation path of sound to each ear and the spectral shaping effects of the listener's pinnae, head and torso. This model is the most realistic albeit the most expensive computationally.

- An extension of stereophonic loudspeaker panning for headphones, creating a phantom image of the source that moves as the azimuth changes. This model is the least realistic but also the least computationally expensive as it does not take any time delays or spectral filtering into account.

- A basic binaural audio model was developed from the application of fundamental acoustic principles on a simplified geometric hearing model. This model can be seen as the middle ground between the two previously mentioned models in terms of the trade-off between realism and computational complexity.

The application can be configured to process the incoming audio streams with any one of the above presentation modes, as well as a monaural audio model.

The audio communications application was integrated with Second Life. Distance attenuation was added to the acoustic model. The virtual world platform provides a more dynamic, conversational environment. Users do not need to "dial" another user that they wish to speak to, they need merely move into audible range, have their conversation and then move away.

Chapter 6

Measurements and Results

The spatial audio communications system developed earlier and the general concept of using spatial audio in electronic voice-based communication will be evaluated in this chapter.

Implementing spatial audio in voice-based telecommunications brings certain benefits, such as higher intelligibility, greater speaker identification and a more immersive experience for users. If has also has certain costs for implementation, in terms of processor usage as well as a possible loss of flexibility. All spatialisation is implemented on the client making for no additional network usage over that usually required for VoIP communication. This chapter will go about measuring these benefits and weighing them against the costs and then evaluating the usefulness of spatial audio in a communications context.

The functioning of the application developed in Chapter 5 on page 50 will first be validated. A website framework will then be developed to conduct human subject experiments. Four psychoacoustic experiments will be performed, the first three evaluating the effect of spatial audio on a listener's ability to identify the active speaker and follow the speech of a target speaker multiple speaker situations. The final psychoacoustic experiment will determine the effect of audio encoding and compression on spatialisation. Finally, the computational costs of each model will be measured.

6.1 Validation of PJSIP Spatialisation

The fidelity of the spatialisation implemented in Section 5.3 on page 54 needs to be validated before the system can be used. The left and right channels are evaluated separately, with the process for each channel being detailed below. A sentence from the Grid corpus, discussed in Section 6.2.1 on page 72, is used as the input. A benchmark signal, $x[n]$, is generated by block convolution of the input by an HRTF pair and normalising the result. A PJSIP call is made, with the source being the same input sample, and spatialised using the same HRTF pair. The call recording functionality of PJSIP gives this signal as $y[n]$. If the spatialisation is functioning correctly then

$$y[n] = \alpha x[n - D] + \omega[n], \tag{6.1}$$

where α is some scaling factor, D some delay and $\omega[n]$ additive noise with $\omega[n] \to 0$. The similarity between the two waveforms can be measured using cross-correlation [115]

$$R_{xy}[l] = \sum_{n=-\infty}^{\infty} x[n+l]y[n], \tag{6.2}$$

where l is the time lag between the signals in samples. The large peak in the cross-correlation shown in Figure 6.1 shows the high degree of correlation between the signals. The value of l for which $R_{xy}[l]$ is a maximum is the time shift D between the signals. The test signal is time-aligned by shifting it D samples and normalised to remove the effect of the scaling factor α, giving $z[n]$. The error function $e[n] = x[n] - z[n]$ is shown to be minimal in relative to $x[n]$ in Figure 6.2 and is most likely the result of quantisation error as the PJSIP frames are stored internally as 16 bit signed integers. The Mean Squared Error (MSE) is $2.84451852537 \times 10^{-5}$ for the left channel and $2.79502308716 \times 10^{-5}$ for the right channel. We can therefore say, with a high degree of certainty, that the PJSIP spatialisation is functioning as should.

6.2 Speech Corpora

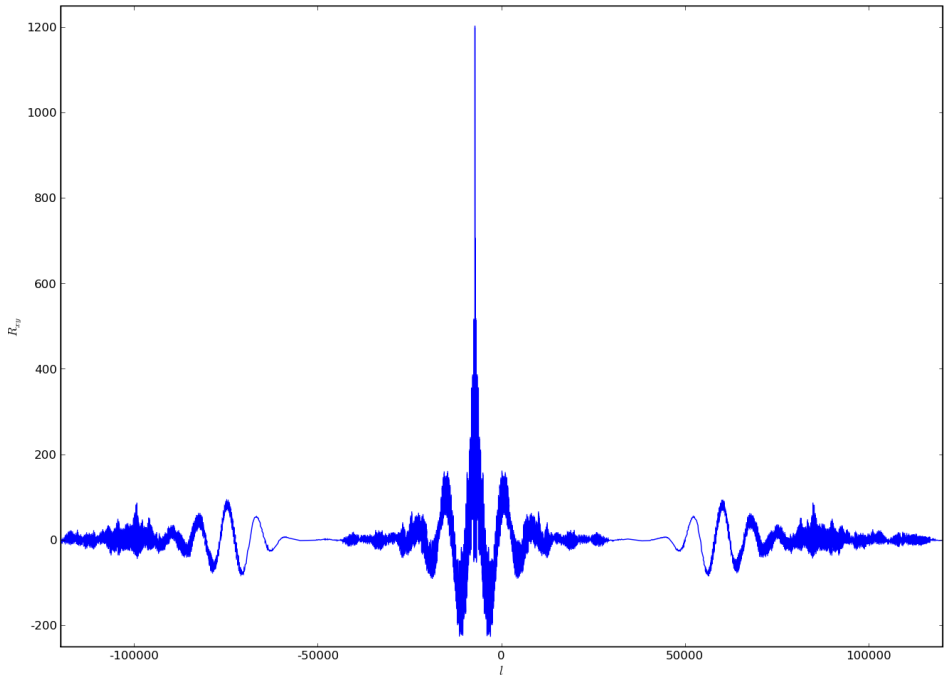
Any audio communication system will require voice recordings for testing and experimentation. In this section we will consider two such speech corpora that will be used as the input audio file for the experiments to be performed in this chapter. The Grid and CMU Arctic corpora will be discussed as well as the pre-processing necessary before they can be used in the experiments.

6.2.1 Grid Speech Corpus

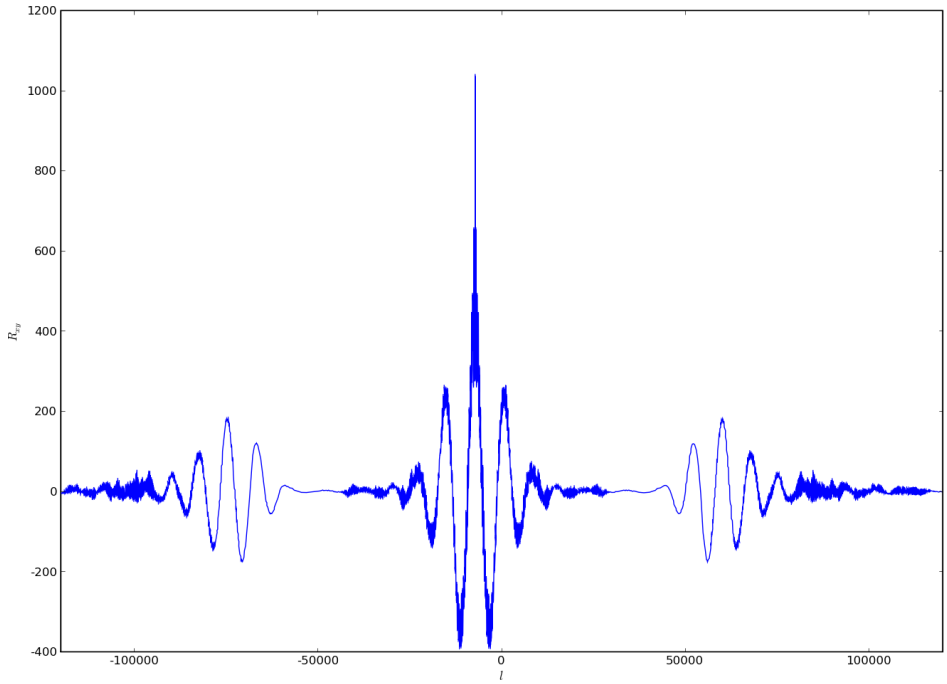
The Grid corpus is a large audiovisual speech corpus [60]. The sentences are of the form *<command><colour><preposition><letter><number><adverb>*, for example "place blue by G4 now" and are provided at a sampling rate of 50 kHz for the raw audio and 25 kHz for audio files that have been endpointed so that the file begins and ends with the begin and end of the utterance. The corpus contains 1000 spoken sentences from each of 34 speakers, eighteen male and sixteen female. All of the samples are under three seconds in duration.

6.2.2 CMU Arctic Speech Corpus

The CMU Arctic speech corpus was designed for speech synthesis research and consists of approximately 1200 phonetically balanced English sentences [98]. The corpus consists of two sets. Set A is the first run of diphone extraction, was recorded in the morning and contains 593 prompts. Set B was recorded in the afternoon and contains 539 prompts. Set A is larger than set B because it contains diphones that only appear once in the corpus. The audio files are sampled at 32 kHz.

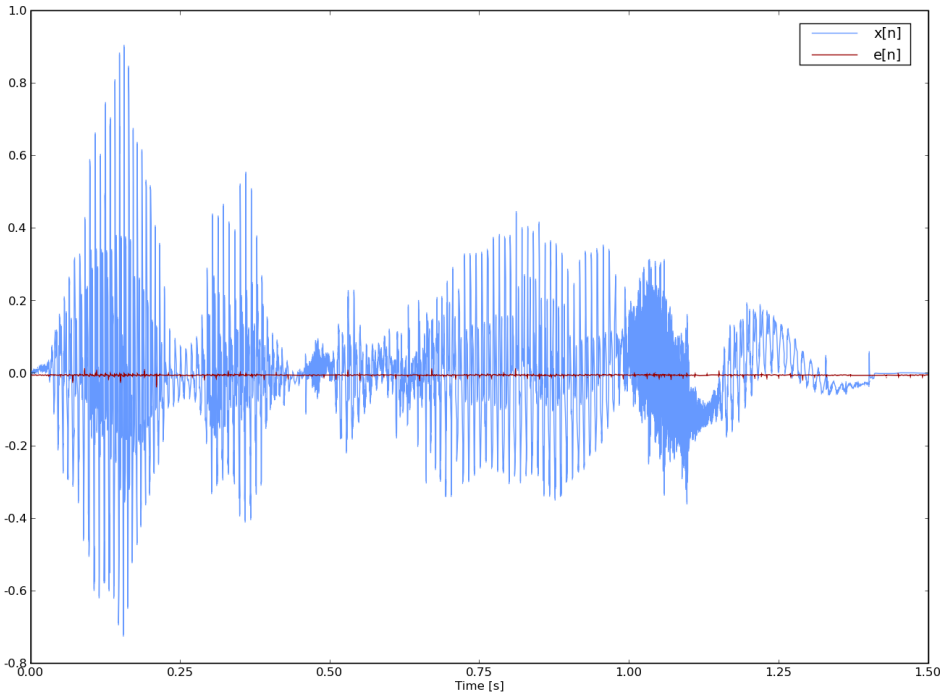


(a) Left channel

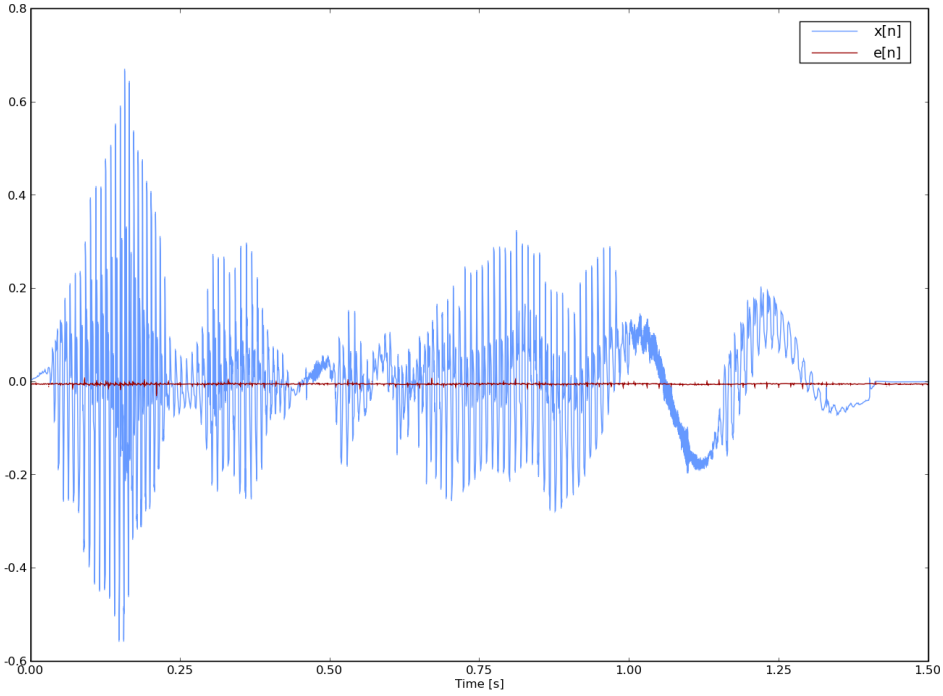


(b) Right channel

Figure 6.1: *Cross-correlation of spatialised sample from PJSIP and block convolution.*



(a) Left channel



(b) Right channel

Figure 6.2: *Error function for PJSIP spatialisation.*

Table 6.1: *Statistics concerning the durations of the audio file sets to be used in speech intelligibility experiments.*

(a) Before removing samples with incorrect length.

Sample set	**Sample count**	**Mean duration [s]**	**Minimum duration [s]**	**Maximum duration [s]**	**Standard deviation [s]**
Grid (subject 3)	1000	2.16	1.67	2.87	0.18
Arctic (bdl)	1131	2.5	0.78	5.55	0.76
Arctic (jmk)	1114	2.6	0.73	5.48	0.78
Arctic (slt)	1132	2.58	0.69	4.89	0.74

(b) After removing samples with incorrect length.

Sample set	**Sample count**	**Mean duration [s]**	**Minimum duration [s]**	**Maximum duration [s]**	**Standard deviation [s]**
Grid (subject 3)	839	2.11	1.67	2.34	0.15
Arctic (bdl)	635	3.04	2.35	5.55	0.51
Arctic (jmk)	675	3.1	2.35	5.48	0.53
Arctic (slt)	696	3.04	2.35	4.89	0.52

6.2.3 Speech Corpora Pre-processing

The speech intelligibility experiments to be discussed in Section 6.9 on page 88 and Section 6.10 on page 93 require the test subject to identify key words spoken by a target speaker in the presence of a number of masking speakers.

Audio samples from talker 3 of the Grid corpus were used as the target speaker. Audio samples from the "bdl", "jmk" and "slt" speakers of the CMU Arctic corpus were used as the first three maskers. All the speakers from the Grid and CMU Arctic corpora are male and speak English as their first language.

For proper masking of the target speaker, it is essential that the all the masking samples are longer in duration that all the target samples. Figure 6.3 shows a Probability Density Function (PDF) of the duration of the files in each audio sample set. Table 6.1(a) shows statistics concerning the duration of the samples in the sets. The cutoff duration is selected to be 2.35 seconds, approximately one standard deviation above the mean of the target set. Target samples shorter than the cutoff and marker samples longer than the cutoff were discarded, Table 6.1(b) shows the statistics of the new sets. The target set is larger than the masker sets, which is acceptable because more variety in the samples that the subjects need to identify is preferable to the other way around. White noise, 2.35 seconds in duration and sampled at 32 kHz, was used as the fourth masker.

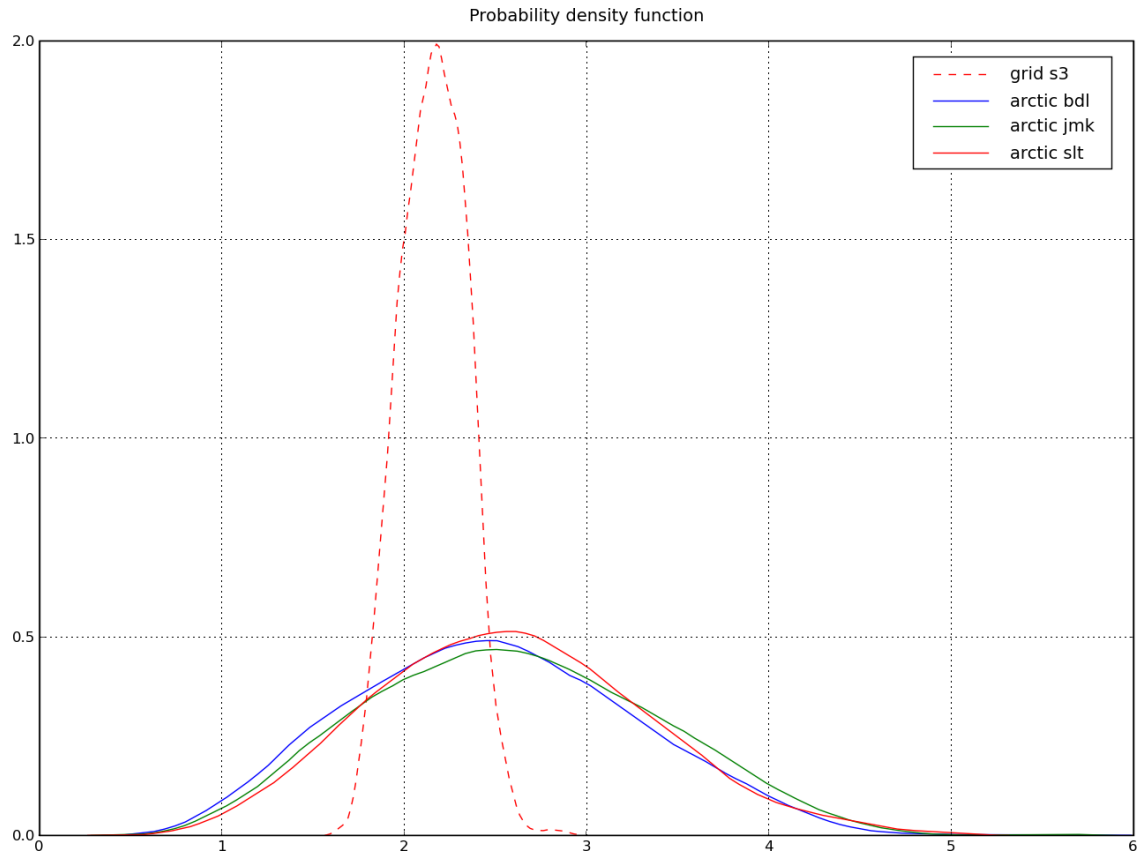


Figure 6.3: *PDF of the duration of the files in each of the sets of audio samples to be used for speech intelligibility experiments.*

6.3 Experimental Framework

Human-subject psychoacoustic experiments will be performed to evaluate the benefit of spatial audio over monaural audio. The subjects will have to listen to audio samples and then either attempt to identify the active speaker for the speaker identification experiments, or the key words of a spoken sentence in the experiments measuring speech intelligibility. To make the experiments easier to administer, they will be performed using a website.

The experiments that will be detailed in Sections 6.8, 6.9 and 6.10 on pages 6.8, 6.9 and 6.10 respectively were administered using a website. This section gives a generic framework that is further customised for the purposes and requirements of each experiment.

6.3.1 Website

The website for the experiments combines elements of HTML, PHP: Hypertext Preprocessor (PHP), Flash and My Structured Query Language (MySQL). The layout of the website is built using HTML. All data processing is done using PHP. The results from each test run are stored in a MySQL database. The audio samples are played using a modified version of the WordPress Audio Player built by Martin Laine [101], a Flash audio player built specifically for use as a WordPress plug-in but capable of being used elsewhere [107]. An audio player built on the Adobe Flash Player platform makes sense due to the high penetration of Adobe Flash Player [32] and the non-standard use of the HTML *object* tag in Microsoft Internet Explorer [57].

At the start of the experiment the subject is redirected to a page with an HTML form asking the subject's name, age and gender as well as giving basic information about the nature of the experiment. The subjects are instructed that headphones are mandatory and a simple check is done by asking the subjects if they are using loudspeakers or headphones. Submitting the form brings up the test page. If the test subject did not specify that they are using headphones, they are redirected to a page instructing them to do this. The transgression is logged in a table in the MySQL database and a link is provided back to the start of the experiment. This method will unfortunately not stop a person dedicated to corrupting the experiment.

The test page instructs the subject about the nature of the target and masking sources and about how to proceed with the experiment. The test page then has an introduction sample that contains only a target sentence in order to familiarise the subject with the target. The five test sentences then follow, with an answer section consisting of radio buttons with options for the colour, letter and number that the subject heard after each audio player instance. The submit button at the bottom of the test page sends the subject to a page that processes the subject's answers and sends them to the MySQL database. The correct answers for each test are read from the seed files and also sent to the database, ensuring that the database contains all the information necessary to calculate the results for the experiment.

The subject should not be allowed to play any sample more than once. Version 1.2.3 of

the WordPress Audio Player was modified to only play the audio sample once each time it is loaded. The refresh detection is done by checking for the existence of a "refresh" cookie before displaying the test page. If the cookie is not found then it is set. If the cookie is found then the subject is redirected to a page that instructs the subject not to refresh the test pages and has a link sending the subject back to the start of the experiment. Both the details page and the process page clear the cookie. All transgressions are logged in the MySQL database. Control questions will also be inserted in the experiment in order to detect cheaters. The measures taken above are still susceptible to exploits from dedicated individuals but are sufficient to hinder casual cheaters. The only failsafe method to ensure that absolutely no cheating occurs would be to administer the experiment under observation.

A flowchart summarising the process of the experiment is shown in Figure 6.4. Variables are passed between the different pages using "hidden" form fields. A screen capture of one of the test samples and answer sections is shown in Figure 6.5. The number of unique users is monitored by placing a cookie on the subject's computer the first time they access the experiment site and checking for the existence of this cookie on subsequent visits.

A MySQL table is created for each experiment that stores all data that might be relevant. Each test run is stored as a new entry in the database. The data stored in each table includes,

- the name, age, gender and IP address of each subject,
- details about the test run that was performed,
- the times that each run was started and completed,
- the answers the subject gave and
- the correct answers.

All the data necessary the process the results for each experiment is included in the database, making it self-contained.

6.4 AMT

Subjects for the experiments will also be recruited using the Amazon Mechanical Turk (AMT) web service.

6.4.1 Description

AMT [2] is a web service from Amazon that provides a crowdsourcing marketplace that gives businesses, referred to as "requesters", access to a scalable and on-demand workforce. Workers do tasks, known as Human Intelligence Tasks (HITs), to earn money, whenever they find convenient. Amazon calls AMT "Artificial Artificial Intelligence". A HIT is a small task that, although fairly easy to accomplish, requires a human to solve. An example of a

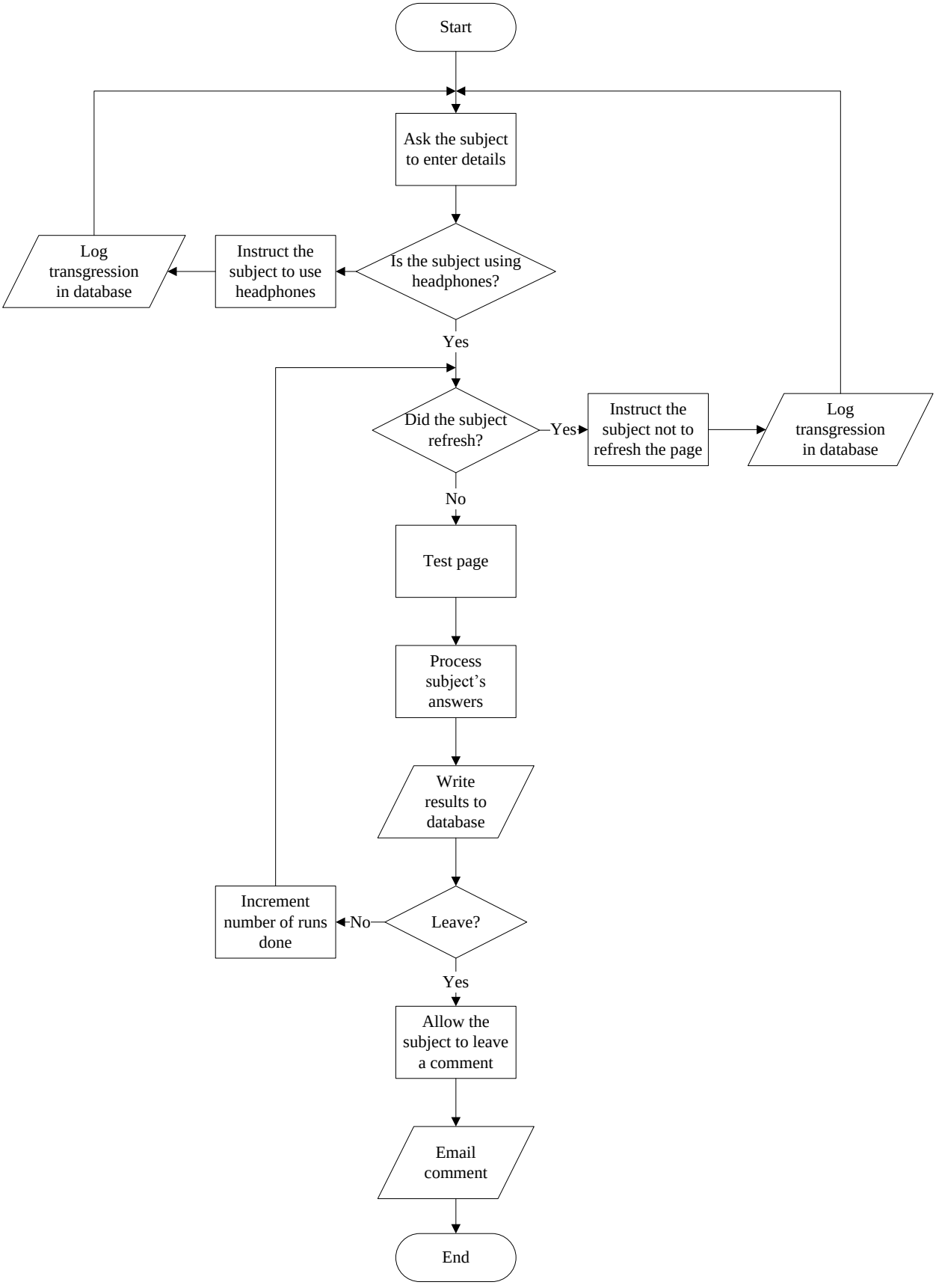


Figure 6.4: *Flowchart summarising how the subject works through the experiment website.*

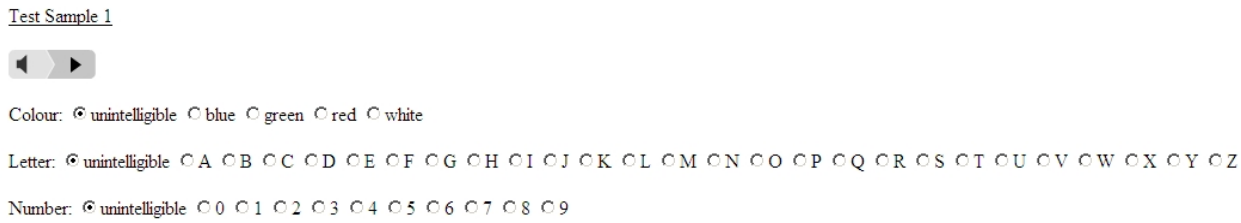


Figure 6.5: *Screen capture of a part of the experiment website showing the audio player for a single test sample along with the radio buttons for the subject's answers.*

HIT would be making comments about the colour scheme of a website, earning the worker $0.05 per website [2].

Although anybody can register to be a worker, a United States bank account and postal address are mandatory for businesses or individuals wishing to register as a requester, making it difficult for a South African to use the AMT service directly. HIT-Builder by DPA Software is a web service that allows people in the rest of the world to make use of AMT through their account. After registration and funding of an account, the HIT-Builder website can be used to generate HITs.

6.5 Psychoacoustic Experiments using AMT

AMT provides an inexpensive platform for performing human-subject experiments, although care needs to be taken to ensure that "casual cheaters" do not corrupt the data collected [97, 87]. Designing the experiment in a manner that does not reward the subject based on their answers will discourage cheating.

Web-based auditory experiments can be done using AMT. It is not possible or practical to embed the experiment websites directly into the HIT page, so the HIT has some introductory text, a link to the experiment website and a single test field for entering a code retrieved upon completion of the experiment. Completing an experiment is considered doing at least ten runs. A modified version of each experiment is created which generates a completion code, which is also stored in a MySQL database. The AMT is required to enter the completion code into the text field provided on the HIT page in order to receive payment. The completion codes are never validated, we believe that the mere possibility of validation is enough to deter casual chancers. The AMT websites store results in a different MySQL table than the regular versions as the AMT data is not considered as dependable. Some of the HITs pay $0.10 and others $0.20.

6.6 Results Processing

Any human-subject experiment where the subjects are not monitored closely will have bad data resulting from participants attempting to cheat the system. These cheaters must be removed from the data pool without biasing the results of the experiment. This is even more important for the data from the experiments conducted using AMT where the high level of anonymity would likely result in more chancers. Subjects that complete the test runs faster than is possible to just listen to all the samples are removed along with those who gave the default answer to every single questions, showing that they did not listen to the samples.

The results from the experiments conducted through normals mean (not through AMT) can be used to find trends that will help in eliminating AMT subjects with suspicious results. Details of exactly how this is done for each experiment will be given in the relevant section.

6.7 Speaker Identification in a Conference Call

We conducted an experiment to evaluate the effect of spatial audio on speaker identification rates for a multiple speaker conversation. The experiment was designed to emulate a conference call, with participants that are unfamiliar to the user, but where it is important to know who is speaking. For example, during a contract negotiation or job interview one would answer questions differently depending on who is asking them. People who do not regularly participate in conference calls might need assistance in identifying the active speaker.

6.7.1 Hypothesis

We hypothesise that spatial audio affords a listener greater ability in identifying an active speaker in a multiple speaker situation than monaural audio.

6.7.2 Experimental Method

We chose to take an off-line approach and use pre-generated audio files instead of testing using the live system to ensure repeatability with different test subjects.

Sentences were chosen from the Grid audiovisual speech corpus that was discussed in Section 6.2.1 on page 72. The 25 kHz endpointed audio files were used for the system after being resampled to 32 kHz to match the sampling rate used in the system. Table 6.2 shows which speakers were chosen and the azimuth of each source. All four of the speakers are male.

Ten sets of audio files were generated, with each set containing four files, each with sentences from only a single speaker. There is no temporal overlap between sentences within a single set. The segments in each set consist of two stages: the introduction stage and the identification stage. The introduction stage has two sentences from each speaker, for a total of eight sentences, in order to allow the subjects to familiarise themselves with the voices and

Table 6.2: *Azimuth and speaker for each of the four sources*

Source	Azimuth	Speaker
1	285°	Grid (subject 1)
2	330°	Grid (subject 2)
3	30°	Grid (subject 3)
4	75°	Grid (subject 5)

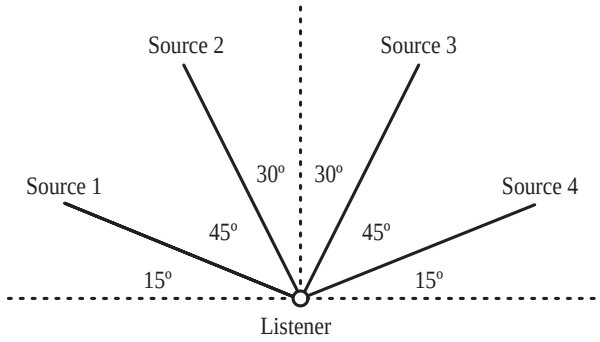
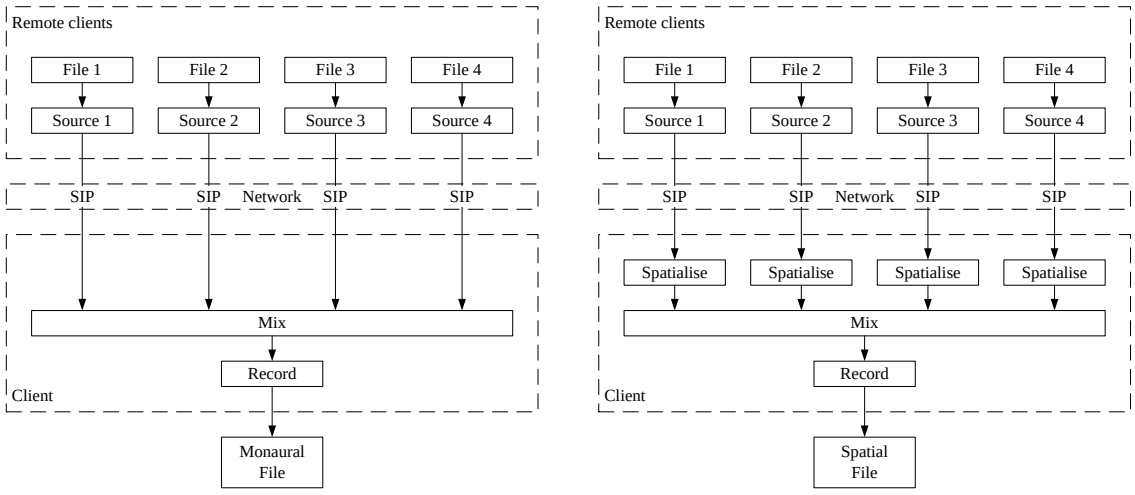


Figure 6.6: *Virtual spatial arrangement of listener and sources in experiment.*

spatial positions of each of the speakers. The identification stage which has four sentences from each speaker in a random order. The subjects will only be required to identify the speakers in the second stage which requires 160 identifications per subject.

The four audio files in each set were used as the input audio source for an instance of the unmodified PJSUA software on four networked computers. The unmodified software with different levels of compression, where the above mentioned experiment had uncompressed audiowas used to demonstrate that the system can spatialize speakers from a standard SIP client. The software on the test computer made a call to all four source computers simultaneously and recorded the resultant conversation in both monaural and spatial arrangements, resulting in two audio files for each of the ten sets. The spatial position of each source is shown in Figure 6.6. Figure 6.7 shows a block diagram summary of this process.

Sixteen subjects with normal hearing were used to conduct the experiment. The subjects were between the ages of 21 and 28, with an average age of 23.7. Thirteen of the subjects were male and three of the subjects were female. A Python script was written to generate a set of files for each test subject in a random order. Each test consisted of five monaural and five spatial audio files, alternating between monaural and spatial, with half of the subjects starting on a monaural file and the other half on a spatial file. Each test subject had to listen to the test files in the order provided and attempt to identify which speaker was active for each of the speech segments. The test subjects were made aware of the spatial arrangement of the sources by showing them Figure 6.6 before beginning the experiment. The subjects were also made aware of the form of the sentences used in the experiment.



(a) Monaural (b) Spatial

Figure 6.7: *Method by which the spatial audio files used in the experiment were created.*

6.7.3 Results

The experiment was conducted in an environment that had low ambient noise levels, typical to that which one would find in an office environment. The audio files were presented to each of the sixteen subjects using high-quality headphones (Sennheiser HD 457, HD 280 Pro, HD 212 Pro, HD 202 and HD 465 headphones were used) to ensure that no distortion was introduced and that the frequency spectrum was not adversely affected.

If a subject made a mistake, either missing a sentence or identifying too many sentences, the data from that particular run was discarded. A total of six mistakes were made, four in a monaural run and two in a spatial run. This left 76 valid monaural runs totalling 1216 valid monaural identifications and 78 valid spatial runs totalling 1248 valid spatial identifications. Figure 6.8 shows the speaker identification rates averaged across all the subjects for each of the monaural and spatial runs, with the propability of guessing correctly included for comparison purposes. On average, the subjects identified the source correctly 43% of the time when the system was in monaural mode and 88% of the time when the system was in spatial mode.

6.8 Speaker Identification in a Multiple Speaker Situation

This experiment was conducted to determine the effect of spatial audio on a subject's ability to recognise a speaker in a situation where any one of multiple speakers could be active at any given time, but not at the same time.

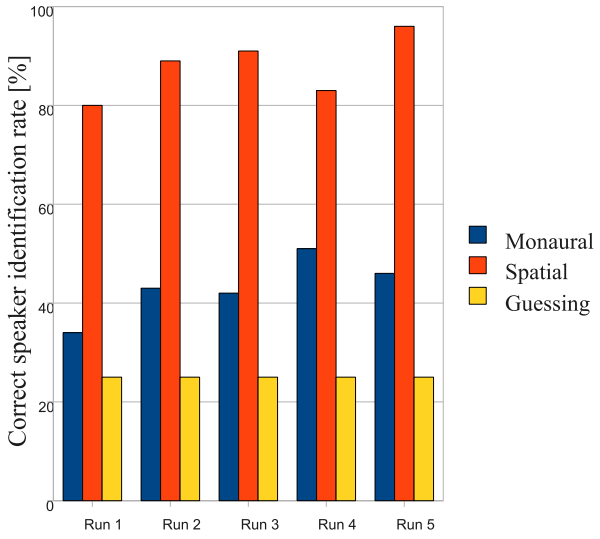


Figure 6.8: *Correct speaker identification rates determined from an experiment with monaural and spatial audio. The probability of guessing correctly is included for comparison purposes.*

6.8.1 Hypothesis

We hypothesise that spatial audio affords a listener greater ability in identifying an active speaker in a multiple speaker situation than monaural audio. HRTF spatialisation, headphone stereo panning and a basic binaural model are compared against monaural audio.

6.8.2 Experimental Method

We chose to administer the experiment through a website over the Internet rather than letting the test subjects work with a live system in a controlled environment. This approach has two advantages over testing with the live system in that it is more more repeatable as well as being more practical to administer to a larger number of test subjects. After the experiment has been set up once it is fully autonomous and can be repeated ad infinitum. The test samples were generated using Python instead of the PJSIP system in order to further automate the process. This was proven to be equivalent to using the live PJSIP system in Section 6.1 on page 71. The samples were generated by implementing the HRTF spatialisation, headphone stereo panning and the binaural model described in Section 5.1 on page 51 in Python.

The test subjects are required to listen to audio samples that contain a single spoken sentence from one of six possible speakers, only one speaker is active in any given sample. In the monaural presentation mode the test subjects will only be able to discern the different speakers based on their voices. In the stereophonic presentation modes the test subjects will be able to use both the speakers' voices and perceived positions to identify them. Figure 6.9 shows the positions of the six sources. Audio files from the Grid corpus, discussed in

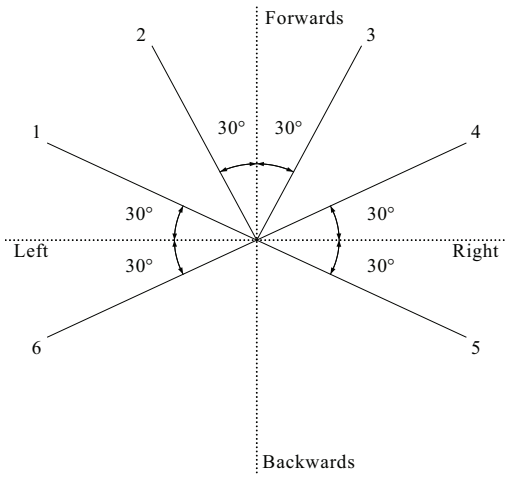


Figure 6.9: *Source positions.*

Table 6.3: *Azimuth and speaker for each of the six sources.*

Source	Azimuth	Speaker
1	300°	Grid (subject 3)
2	330°	Grid (subject 5)
3	30°	Grid (subject 6)
4	60°	Grid (subject 10)
5	120°	Grid (subject 12)
6	240°	Grid (subject 13)

Section 6.2.1 on page 72, were used as the audio sources. Table 6.3 shows which speakers were chosen and the azimuth of each source. All six of the speakers are male.

The experiment has a hundred possible test runs, with each test run having five audio samples. The audio samples that the subjects will listen are generated in two steps. The first step creates a set of seed files that describe each test run. Each seed file has a line for each of the five samples in the test, containing the number of a randomly selected source, the azimuth of the chosen source and the filename of a randomly selected audio sample from the chosen source. An example seed file is shown in Listing 6.1. An additional "introduction" seed file with six lines, one for each of the six sources, in order is also generated. This seed file will be used to generate of set of introduction samples to familiarise the test subjects with the voice of each speaker.

The second step takes each of the 100 test seed files and generates a set of test samples for each of the four presentation modes, making a total of 400 tests. Each test consists of a directory with five audio samples, containing a single speaker uttering a single sentence. Each test sample is generated by taking the appropriate audio file as specified by the seed file (resampled to 32 kHz) and processing it with the algorithm for the current presentation mode at the azimuth specified in the seed file. Figure 6.10 shows the process for generating a

Listing 6.1: *Example seed file.*

```
3 30 sbaazs
6 240 swit4p
2 330 lgwg1a
6 240 lwij9a
5 120 brwf8a
```

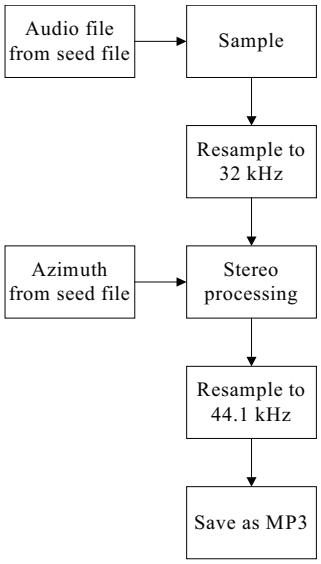


Figure 6.10: *Sample generation process for a single sample.*

single test sample. The stereo processing block can be monaural, spatial, panning or binaural processing, depending on the current presentation mode. The output audio samples are saved as 44.1 kHz, 320 kbps MP3 files as required by the Flash audio player. A set of audio files corresponding to the introduction seed file is also generated in the same fashion.

The experiment is administered using a modified version of the experiment website framework discussed in Section 6.3 on page 77. Following the details page, the introduction page provides further information on the experiment including a figure with source positions, an example test and answer section and audio player with a monaural sample for each speaker to familiarise the subject. Figure 6.11 shows a test and answer section. The subjects are made aware that the sound in some of the audio samples will appear to come from a certain direction and that they need to use the speaker's voice and position to identify them. The introduction page leads on to the test page which also has audio players introducing each speaker. The five test and answer sections then follow, after this the website adheres to the framework.

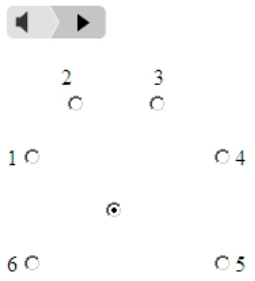


Figure 6.11: *Screen capture of a part of the experiment website showing the audio player for a single test sample along with the radio buttons for the subject's answers.*

6.8.3 Results

For the experiment conducted through normal means, 62 subjects took part in the experiment and 45 subjects completed at least five test runs. Data from subjects with fewer than five test runs was dropped. The subjects were between the ages of 19 and 45, with an average age of 24.19 years. Males accounted for 82.26% of the subjects and females for 17.74%.

Asking AMT workers for identity information like their names is a contravention of the AMT terms of service. This makes determining the total number of subjects impossible, as two subjects of the same age and gender will appear to be one if names cannot be used to identify them. Cookies are not completely reliable for identifying users as many people turn them off. However, 236 people completed the fifteen runs required for payment meaning that there are at least as many subjects. Data from subjects with fewer than ten test runs was dropped. The subjects were between the ages of 18 and 56, with an average age of 30.83 years. Males accounted for 62.07% of the subjects and females for 37.93%. The average age and gender distribution are rough approximations, again due to it being impossible to know the exact number of unique subjects.

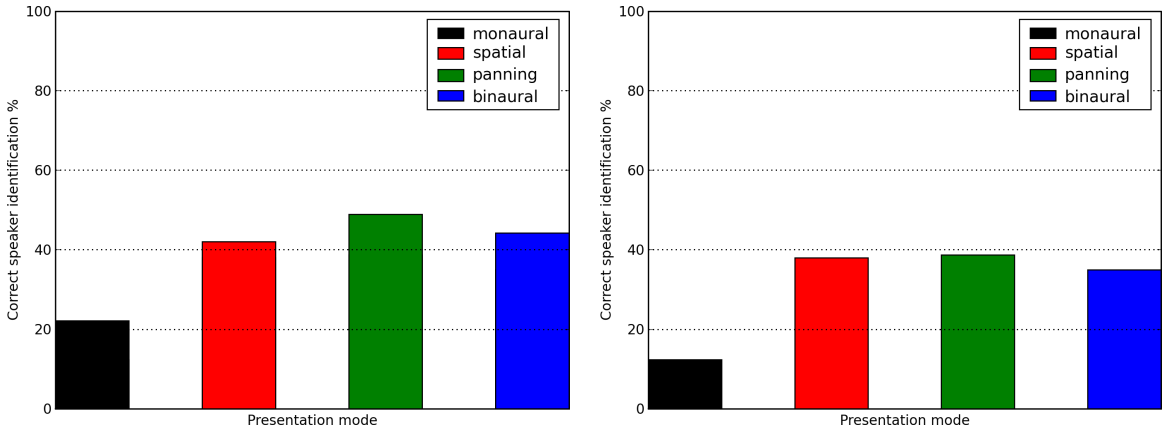
The data collected from the AMT experiments needed to be cleaned. All data from subjects that completed each test run in a time that is less than required to listen to all the audio samples back to back is dropped. All subjects that make side-to-side error in any of the stereo presentation modes, something that can only occur if the subject is cheating, has severe hearing problems or has their headphones on the wrong way around, is dropped. No individual tests results were discarded, only results from any subjects deemed suspicious. We believe that this does not compromise the integrity of the experiment.

Table 6.4 shows the number of test run in the collected data, both before and after cleaning. Each test consists of one speaker direction identification. The AMT experiment collected a large amount of data but test results from a large number of subjects needed to be dropped, leaving only 17.4% of the total data collected as valid. The normal experiment data was much more reliable, with 93.27% of the tests being used.

Figure 6.12 shows the correct speaker identification rates for the experiment, both through the normals means and through AMT. The spatial presentation modes greatly surpass

Table 6.4: *The number of test results collected for each presentation mode.*

	Monaural	HRTF	Panning	Binaural
Normal (before cleaning)	965	915	925	985
Normal (after cleaning)	885	875	860	915
AMT (before cleaning)	3480	3200	3480	3315
AMT (after cleaning)	600	495	680	570



(a) Normal experiment

(b) AMT experiment

Figure 6.12: *Speaker identification rates for multiple speaker situations.*

the monaural presentation, with the normal HRTF spatial (42.06%) and binaural (44.26%) modes achieving scores less than twice the monaural mode (22.15%). The headphone panning model performs the best out of the three spatial models. The basic binaural model and HRTF spatialisation show similar improvements of 40.02% and 39.58% respectively when averaging across both experiments.

The overall scores are significantly lower than those of the earlier experiment described in Section 6.7 on page 81. This can largely be attributed to the larges number of possible sources positions, six versus four in the earlier experiment. Evaluating four different spatialisation methods could also make localisation more difficult for the subject, as the exact positional cues vary between the methods.

6.9 Speech Intelligibility in a Multiple Speaker Situation

This experiment was conducted to determine the effect that localised audio has on speech intelligibility in a multiple speaker situation where more than one speaker is active at any given time. The aim is to emulate the cocktail party effect in which a person has to follow the speech of a target speaker in the presence of masking speakers.

Listing 6.2: *Example seed file for a scenario 6 test run.*

```
60 150 315
sgbp4p
prbp8p arctic_a0531 arctic_a0116
prbj2n arctic_b0533 arctic_a0563
bgwo1s arctic_b0279 arctic_b0494
swiu4p arctic_a0582 arctic_a0087
swib3a arctic_a0191 arctic_b0165
```

6.9.1 Hypothesis

We hypothesise that spatial audio provides greater speech intelligibility than monaural audio in a situation with more than one concurrent speaker. HRTF spatialisation, headphone stereo panning and a basic binaural model are compared against monaural audio.

6.9.2 Experimental Method

The experiment is conducted in the same manner as the one detailed in Section 6.8 on page 83 and only the differences will be discussed here.

The target and masking speakers are as specified in Section 6.2.3 on page 75. Table B.1 on page 128 shows the azimuth positions of the target and masker sources for each of the sixteen different scenarios, where N is the number of maskers. Figures B.1, B.2, B.3 and B.4 on pages 124, 125, 126 and 127 show the target and maskers position, where T designates a target, M a speech masker and W a noise masker.

The audio samples that the subjects will listen to are generated in two steps. The first step creates a set of seed files that describe each of the test arrangements. Each seed file consists of a line containing the azimuth values of the $N+1$ sources, where N is the number of maskers in the scenario. The seed file then has a line containing a target filename, following this are five lines containing the $N + 1$ filenames for the target and maskers. An example seed file for a scenario 6 test run is shown in Listing 6.2. All target and maskers files are randomly selected from the audio file sets for each source. Each of the sixteen scenarios is repeated five times to provide a greater number of answer possibilities, for a total of eighty seed files.

The second step takes each of the eighty seed files and generates a set of test samples for each of the four presentation modes, making a total of 320 tests. Each test consists of a directory with six audio samples, the first containing only the target speaker to familiarise the subject. The other five files contain the target speaker in the presence of the masking speakers. Each test sample is generated by taking all the source samples for that case (resampled to 32 kHz), processing them with the algorithm for the current presentation mode and mixing them together as a single stereophonic audio stream (or monaural stream

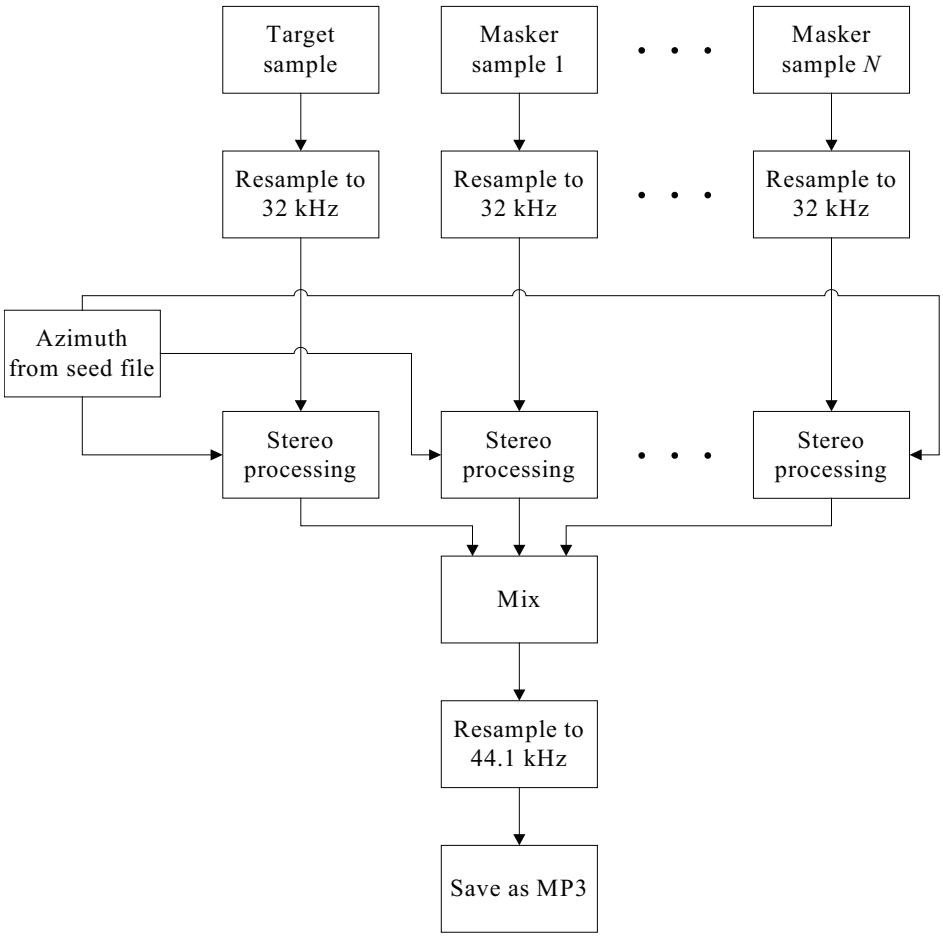


Figure 6.13: *Sample generation process for a single sample, N is the number of maskers in the current positional scenario.*

for the monaural presentation mode). Figure 6.13 shows the process for generating a single test sample. The stereo processing block can be monaural, spatial, panning or binaural processing, depending on the current presentation mode. N is the number of maskers in the current positional scenario and varies between one and four. The azimuth for each of the $N + 1$ sources is taken from the first line in the seed file of each test run. The output audio samples are saved as 44.1 kHz, 320 kbps MP3 files as required by the Flash audio player.

The experiment is administered using the website framework discussed in Section 6.3 on page 77.

6.9.3 Results

For the experiment conducted through normal means, 78 subjects took part in the experiment and 53 subjects completed at least five test runs. Data from subjects with fewer than five test runs was dropped. The subjects were between the ages of 19 and 56, with an average age of 25.94 years. Males accounted for 84.62% of the subjects and females for 15.38%.

For the AMT experiment, 230 people completed the fifteen runs required for payment

Table 6.5: *The number of test results collected for each presentation mode.*

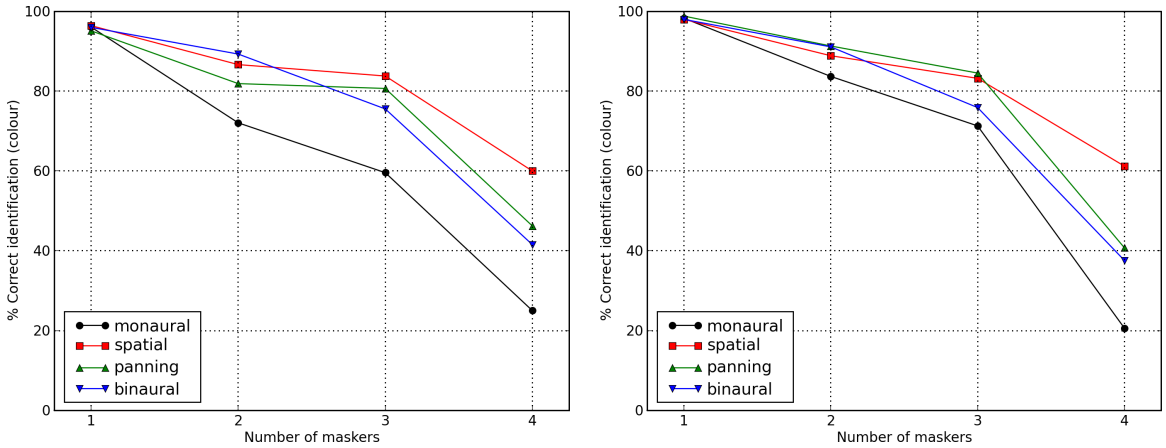
	Monaural	**HRTF**	**Panning**	**Binaural**
Normal (before cleaning)	1070	1060	860	910
Normal (after cleaning)	980	950	765	825
AMT (before cleaning)	3230	3285	3025	3035
AMT (after cleaning)	590	655	565	695

meaning that there are at least as many subjects. Data from subjects with fewer than ten test runs was dropped.

The performance for the one masker situation, which is very easy, was used as a control. The normal experimental subjects had an average colour and number identification rate of no less than 92.9% across all presentation modes. Any subjects that have an average colour and number identification rate for the one masker scenario of less than 40% were dropped. The subjects were between the ages of 18 and 60, with an average age of 34.3 years. Males accounted for 43.33% of the subjects and females for 56.67%. The average age and gender distribution are rough approximations, again due to it being impossible to know the exact number of unique subjects.

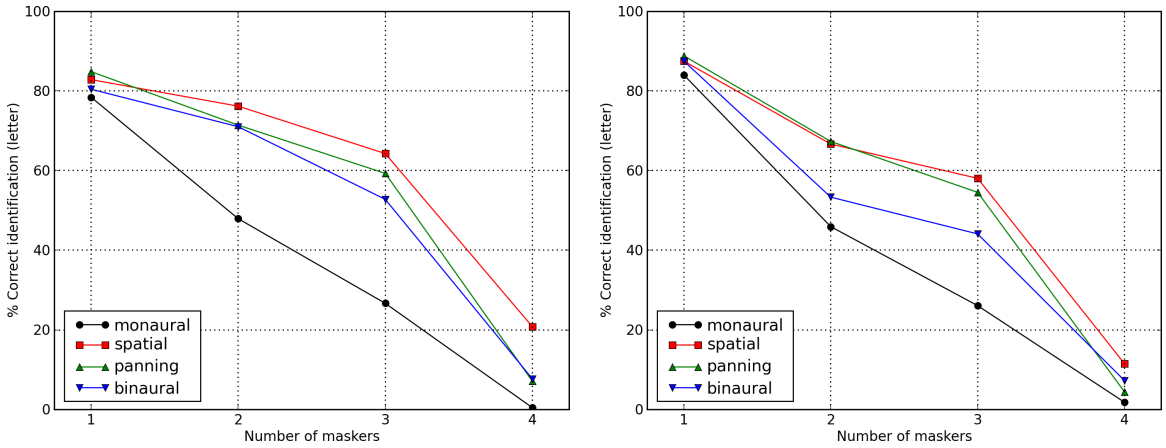
Table 6.5 shows the number of test run in the collected data, both before and after cleaning. Each test consists of three identifications, one each of the colour, letter and number key words. The AMT experiment collected a large amount of data but test results from a large number of subjects needed to be dropped, leaving only 19.92% of the total data collected as valid. The normal experiment data was much more reliable, with 90.26% of the tests being used.

The intelligibility rates of the colour, letter and number key words as a function of the number of masking speakers are shown in Figures 6.14, 6.15 and 6.16 respectively. The results for both the experiment conducted using AMT and the experiment conducted through normal means are given in each figure. On average the intelligibility scores are the highest for the colour key word and the lowest for the letter key word which is exactly as expected because of the number of possible choices. When a single masking speaker is active all presentations modes perform well, as expected for such an easy scenario. As the number of maskers increases, the intelligibility of all modes decreases, but the three spatial modes less so than the monaural mode. The HRTF spatialisation outperforms all the other modes when four speakers are active. Figures 6.14(a), 6.16(a) and 6.16(a) show HRTF spatialisation providing greater intelligibility with three or four masking sources than monaural audio does with one less masker.



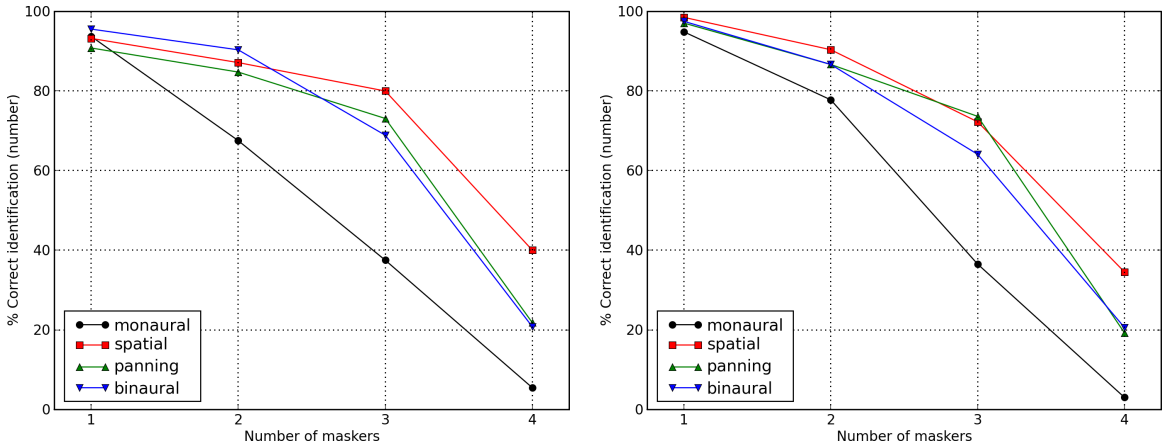
(a) Normal experiment (b) AMT experiment

Figure 6.14: *Intelligibility rates for multiple speaker situations (colour).*



(a) Normal experiment (b) AMT experiment

Figure 6.15: *Intelligibility rates for multiple speaker situations (letter).*



(a) Normal experiment (b) AMT experiment

Figure 6.16: *Intelligibility rates for multiple speaker situations (number).*

6.10 Effect of Audio Encoding and Compression on Spatialisation

Bandwidth limitations make sending raw audio over the network impractical, except in situations where the network in use is a high speed LAN. If the system is to be used over the Internet the audio packets will need to be encoded in order to save bandwidth. A spatial audio telephony system that only works with uncompressed audio is not very practical to use in the real world.

6.10.1 Aim

The aim of this experiment is to determine the effect of different levels of encoding and compression on the effectiveness of spatial audio. This is done to determine if spatial audio will function as well as monaural audio does with audio streams that have been compressed and encoded.

6.10.2 Experimental Method

This experiment is very similar to the experiment discussed in Section 6.9 on page 88 except that the number of maskers and positional arrangement remain constant and that the quality audio samples used to generate the test files varies. The quality of the audio samples from the speech corpora is degraded by encoding and decoding the samples with a lossy audio codec. Only the differences from the aforementioned experiment will be discussed here.

To simulate the effects of encoding and compressing a live audio stream, the audio samples from the speech corpora will be encoded with *speexenc* (and subsequently decoded with *speexdec*) before generating the test samples. The experiment will ascertain the effect of bitrate and bandwidth on the speech intelligibility of spatial audio, relative to that of monaural audio. This experiment will use a modified version of the website-based experimental platform developed in Section 6.3 on page 77.

The rate of decline in speech intelligibility for spatial audio, as a function of bitrate and bandwidth, will be compared to monaural audio. The target and masking speakers are as specified in Section 6.2.3 on page 75. The positions of the sources is shown in Figure 6.17, where T designates a target and M a speech masker.

The Speex codec, discussed in Section 4.4.1 on page 48, was used to encode the audio samples because of its open-source nature. Speex does not perform as well as closed-source codecs due to licensing issues prohibiting the application of certain techniques, requiring higher bit rates to achieve the same Mean Opinion Score (MOS) values as the Adaptive Multi-Rate (AMR) and Adaptive Multi-Rate Wideband (AMR-WB) codecs [116]. The experiment evaluates the relative differences in performance of the presentation modes after being compressed, making the absolute performance of the codec unimportant and the open-source nature of Speex makes it better for experimentation than the closed-source AMR and

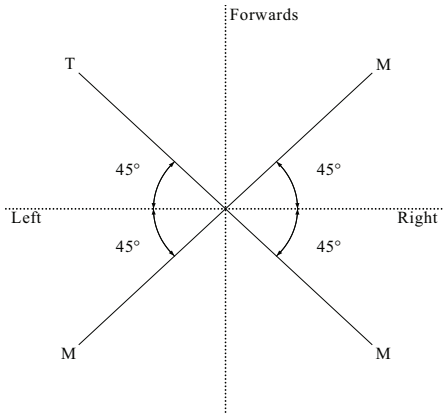


Figure 6.17: *Source positions for scenarios with two maskers, where T designates a target and M a speech masker.*

AMR-WB codecs.

Although the Speex codec offers a number of different bit rates, evaluating the increase in speech intelligibility of each would result in a unnecessarily high number of test cases. The number of test cases can be decreased by only evaluating bit rates that give a large difference in perceived quality. According to a subjective evaluation by the author, an increase in bit rate at the higher rates contributes a much smaller increase in speech intelligibility than an equivalent increase at the lower rates [88]. An experiment by Anssi Rämö and Henri Toukomaa also shows a greater increase in MOS values for an increase in bit rate at lower bit rates than at higher ones [116]. To assist in choosing which bit rates to evaluate, the MSE was calculated for pairs of consecutive bit rates using,

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N} e_i^2, \tag{6.3}$$

where e_i is the difference between the resultant output signals from a reference signal being encoded at two consecutive bit rates and N the length in samples of the reference signal. Table 6.6 shows the MSE between consecutive bit rates using a speech signal as reference. Version 1.1.12 (compiled May 7, 2008) of the codec was used.

Not all of the available bit rates for each bandwidth mode were evaluated in the experiment, only the bit rate pairs that from Table 6.6 that gave $MSE \geq 0.001$ as well as the highest and lowest available bit rates for each mode. Table 6.7 shows the bit rates that were chosen. This gives a total of eighteen different quality combinations, six for each of the three sampling rates.

The audio samples that the subjects will listen to are generated in two steps. The first step creates a set of seed files that describe each of the test arrangements. Each seed file consists of a line containing the quality parameters for the test case, the bandwidth and bit rate. The seed file then has a line containing a target filename, following this are five lines containing the filenames for the target and three maskers. An example seed file for a

Table 6.6: *MSE of consecutive bit rate pairs for Speex codec.*

Bandwidth	**Bit rate 1**	**Bit rate 2**	**MSE**
8 kHz	2,150 bps	3,950 bps	0.0358254767
8 kHz	3,950 bps	5,950 bps	0.0101565771
8 kHz	5,950 bps	8,000 bps	0.0017101937
8 kHz	8,000 bps	11,000 bps	0.0259890603
8 kHz	11,000 bps	15,000 bps	0.0006358961
8 kHz	15,000 bps	18,200 bps	0.0003169583
8 kHz	18,200 bps	24,600 bps	0.0001760196
16 kHz	3,950 bps	5,750 bps	0.0337654681
16 kHz	5,750 bps	7,750 bps	0.0102933004
16 kHz	7,750 bps	9,800 bps	0.0024642114
16 kHz	9,800 bps	12,800 bps	0.0288206294
16 kHz	12,800 bps	16,800 bps	0.0009520490
16 kHz	16,800 bps	20,600 bps	0.0001321288
16 kHz	20,600 bps	23,800 bps	0.0004503355
16 kHz	23,800 bps	27,800 bps	0.0000464394
16 kHz	27,800 bps	34,200 bps	0.0002605977
16 kHz	34,200 bps	42,200 bps	0.0000127436
32 kHz	4,150 bps	7,750 bps	0.0391912591
32 kHz	7,750 bps	9,550 bps	0.0114259462
32 kHz	9,550 bps	11,600 bps	0.0279762791
32 kHz	11,600 bps	14,600 bps	0.0018548806
32 kHz	14,600 bps	18,600 bps	0.0009840550
32 kHz	18,600 bps	22,400 bps	0.0002670490
32 kHz	22,400 bps	25,600 bps	0.0004869111
32 kHz	25,600 bps	29,600 bps	0.0001003703
32 kHz	29,600 bps	36,000 bps	0.0002740306
32 kHz	36,000 bps	44,000 bps	0.0000391483

Table 6.7: *Speex codec bit rates to be used in experiment.*

Narrowband (8 kHz)	Wideband (16 kHz)	Ultra-wideband (32 kHz)
2,150 bps	3,950 bps	4,150 bps
3,950 bps	5,750 bps	7,550 bps
5,950 bps	7,750 bps	9,550 bps
8,000 bps	9,800 bps	11,600 bps
11,000 bps	12,800 bps	14,600 bps
24,600 bps	42,200 bps	44,000 bps

Listing 6.3: *Example seed file for a 16 kHz 12.8 kbps test run.*

```
16000hz 12800bps
pwij2p
bbbz9a arctic_a0391 arctic_b0256 arctic_a0493
pbbv5s arctic_b0411 arctic_a0546 arctic_a0199
lrak6p arctic_a0538 arctic_a0358 arctic_a0371
srab1a arctic_b0426 arctic_a0219 arctic_a0380
lgwgzn arctic_a0166 arctic_a0382 arctic_a0395
```

16 kHz (wideband), 12.8 kbps test run is shown in Listing 6.3. All target and maskers files are randomly selected from the audio file sets for each source. Each of the eighteen quality combinations is repeated five times to provide a greater number of answer possibilities, for a total of ninety seed files.

The second step takes each of the ninety seed files and generates a set of test samples for each of the four presentation modes, making a total of 360 tests. Each test consists of a directory with six audio samples, the first containing only the target speaker to familiarise the subject. The other five files contain the target speaker in the presence of the masking speakers. Each test sample is generated by taking all the source samples for that case (resampled to 32 kHz), processing them with the algorithm for the current presentation mode and mixing them together as a single stereophonic audio stream (or monaural stream for the monaural presentation mode). Figure 6.18 shows the process for generating a single test sample. The stereo processing block can be monaural, spatial, panning or binaural processing, depending on the current presentation mode. There are three maskers in the positional scenario, shown in 6.17. The output audio samples are saved as 44.1 kHz, 320 kbps MP3 files as required by the Flash audio player.

The experiment is administered using the website framework discussed in Section 6.3 on page 77.

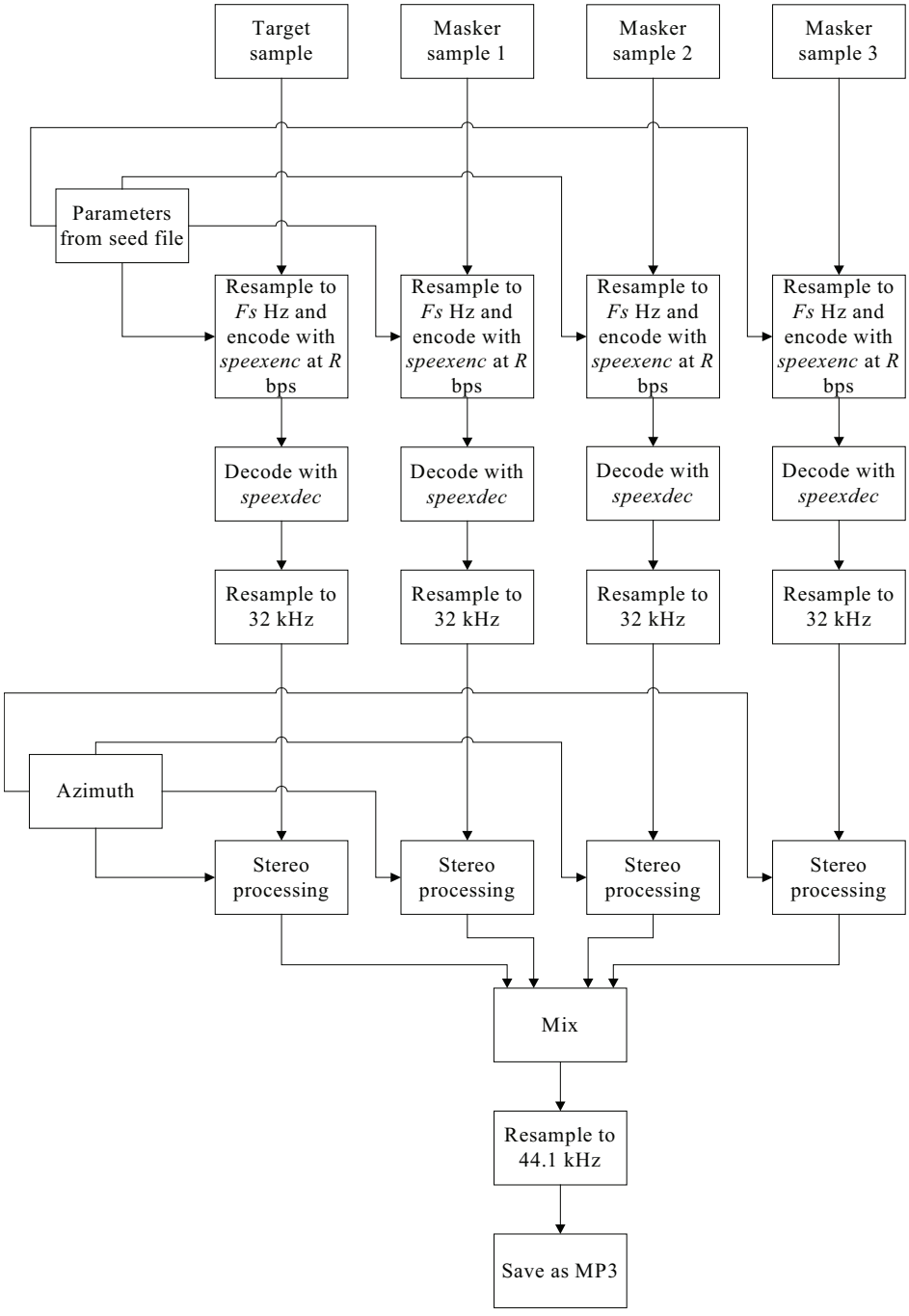


Figure 6.18: *Sample generation process for a single sample, Fs and R are the bandwidth and bit rate of the current test case.*

Table 6.8: *The number of test results collected for each presentation mode.*

	Monaural	HRTF	Panning	Binaural
Normal (before cleaning)	625	630	610	615
Normal (after cleaning)	610	585	590	590

6.10.3 Results

Data acquired through the AMT experiment was not used for lack of a reliable means to screen for invalid results. For the experiment conducted through normal means, 38 subjects took part in the experiment and 34 subjects completed at least five test runs. Data from subjects with fewer than five test runs was dropped. The subjects were between the ages of 17 and 39, with an average age of 23.74 years. Males accounted for 85.35% of the subjects and females for 17.65%.

Table 6.8 shows the number of test runs in the collected data, both before and after cleaning. Each test consists of three identifications, one each of the colour, letter and number key words. The normal experiment data was very reliable, with 95.77% of the tests being used. The average intelligibility rates of the three key words was used for all results.

Figure 6.19 shows the decrease in average word intelligibility, relative to that of monaural audio, as the bit rate decreases from the maximum for the three spatial presentation modes at 16 kHz bandwidth. The absolute intelligibility values are not important, only the relative rate of decline thereof. The values for the monaural mode are subtracted from each of the spatial modes and the value at the maximum bit rate is then subtracted from this. Therefore, any values below zero indicate conditions in which the intelligibility of the presentation mode specified suffers more greatly from a decrease in bit rate than the monaural audio model would. The binaural model does not handle low bit rates very well. The HRTF spatialisation and panning models have decreased performance at 12.8 kbps, but function at least as well as monaural audio everywhere else. Some of this variance could be attributed to the experimental nature of the Speex codec, which is not yet a mature project.

Figure 6.20 shows the intelligibility of each presentation mode at the maximum bit rate as a function of the bandwidth. Monaural audio does not receive any benefit from an increase in bandwidth while the intelligibility of the three spatial modes increase as the bandwidth increases. As the different bandwidth modes do not share the same bit rates, the intelligibility rate relative to the average bit rate provides a better indication an increase in bandwidth provides a similar proportional increase in performance. This is shown in Figure 6.21. A decreasing function means that the performance does not increase in the same proportion that the bit rate increases. The only mode does comes close to maintaining a unity performance to bit rate ratio is the HRTF spatial model when going from 16 to 32 kHz. The spatial modes perform especially poorly when going from 8 to 16 kHz.

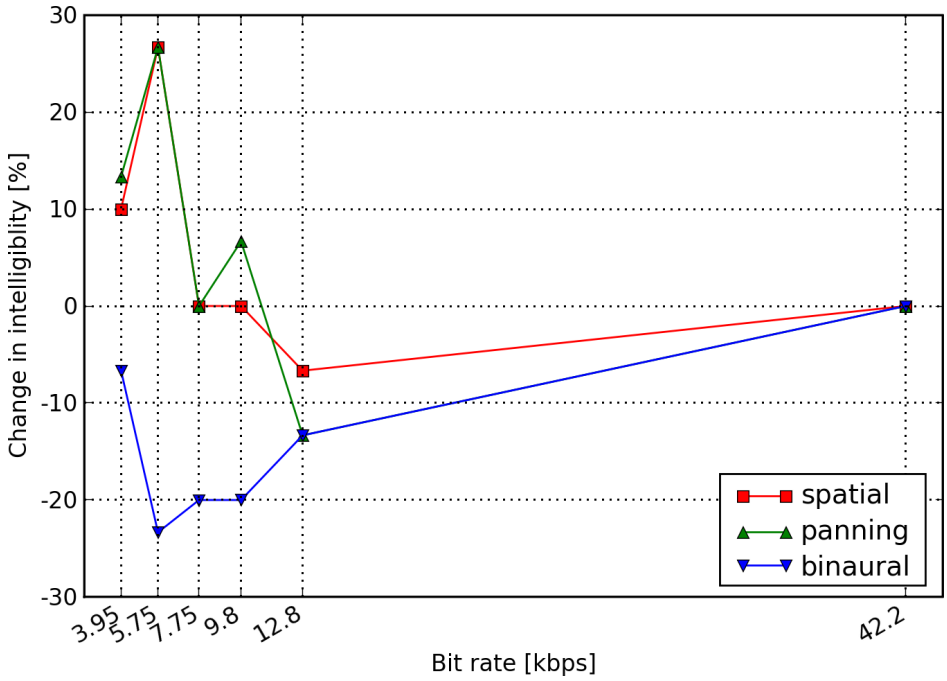


Figure 6.19: *Change in average speech intelligibility relative to monaural audio.*

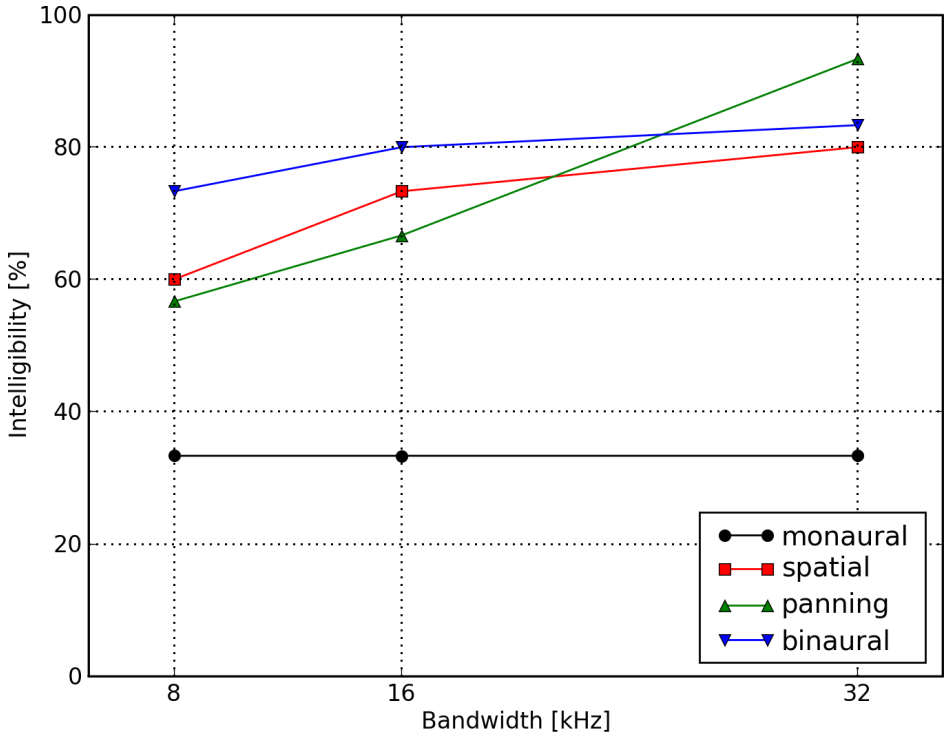


Figure 6.20: *Intelligibility at maximum bit rate for each bandwidth mode.*

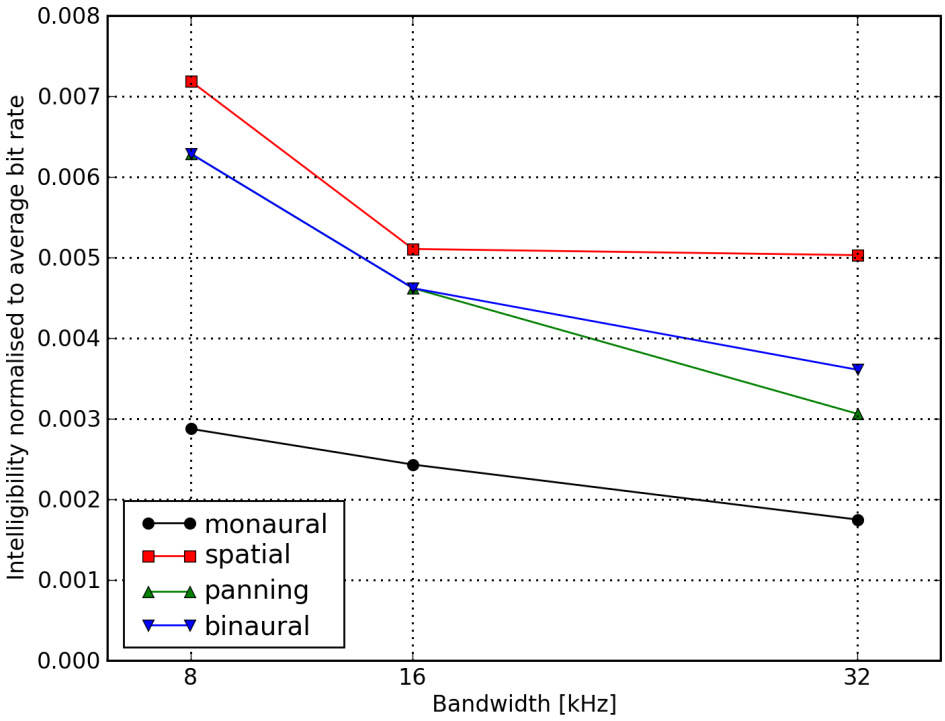


Figure 6.21: *Intelligibility relative to average bit rate for each bandwidth mode.*

6.11 System Benchmarking

The spatialisation algorithms implemented in the application designed in Section 5.3 on page 54 will place an additional computational burden on the system. This burden needs to be quantified in order to ascertain the feasibility of using spatialisation in a real-time audio communications platform.

6.11.1 Aim

The aim of this experiment is to determine the change in system resource usage because of the implementation of spatialisation in PJSIP as a function of the number of calls. This experiment is not a detailed performance analysis of the various audio model implementations, but merely a broad comparison of the processing costs of each model and not too much will be inferred from the results.

6.11.2 Experimental Method

The program needs to be profiled to determine the additional costs of the HRTF, panning and binaural spatialisation models over the existing monaural one. The audio frame processing functions are clocked on 20 millisecond cycles, making the determination of the total execution time of the program irrelevant. Two software profiling packages available on Linux are Valgrind [24] and GNU Profiler (gprof) [4]. The sampling process used in gprof does not

work very well with a program that spends a lot of time in system calls and can return unexpected results [105]. Valgrind is a simulation-based profiling package and does not have such a problem and will therefore be used for performing this experiment. Running a program in the Valgrind simulation environment does unfortunately make it run around 50 times slower. No changes need to be made to the program source code, the binary is run using the Callgrind tool in the Valgrind environment and generates a Callgrind output file for later processing. Callgrind measures function calls and costs. Valgrind measures the number of instructions performed by the processor, while not directly proportional to the time taken, provides a good indication of something that is otherwise very difficult to simulate [11]. For each component that was profiled the inclusive cost, which is the self cost plus the cost of all callees, was measured.

The PJSUA program is modified to run autonomously. The user interface was overridden so that the execution of the program can be scripted without requiring any form of user input. Upon start the program calls all SIP URIs contained in a contacts file. After the call duration specified in the configuration file has elapsed, each call hangs up the program exits after the last call has been completed.

The experiment is conducted using a Python script that edited the start-up options, started a remote unmodified PJSUA client that was to be called, started the modified PJSUA client under Valgrind and stored the resulting Callgrind output file. In this way the experiment can easily be automated to collect data for a number of different call scenarios, a process that would be extremely tedious to do manually. The process is repeated for each of the four presentation modes for the number of concurrent calls from 0 to 32. PJSIP has a hard limit of 32 active calls, which is also the point at which audio quality on the test machine starts to degrade.

6.11.3 Results

The total instruction cost of the program as a function of the number of active calls is shown in Figure 6.22. The first call is the most costly due to fixed costs, after which the cost has a somewhat linear relationship to the number of calls. The panning model has almost exactly the same instruction cost as the monaural one, due to its low computational complexity, which is negligible compared to normal costs incurred when making calls.

The instruction cost per call as a function of the number of active calls is shown in Figure 6.23. The average cost per call in the relatively linear region was calculated for each of the spatial presentation modes, relative to monaural audio. The HRTF spatialisation mode costs 9.03 times as much as the monaural mode, the panning mode 1.03 times and the binaural mode 6.86 times.

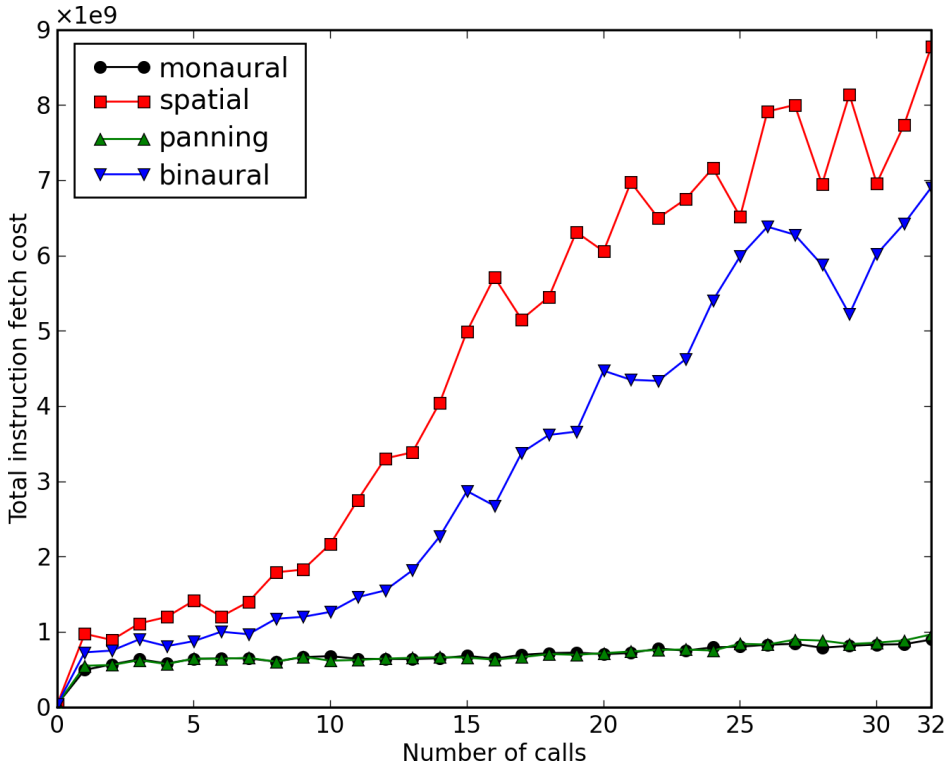


Figure 6.22: *Total instruction cost for each presentation mode.*

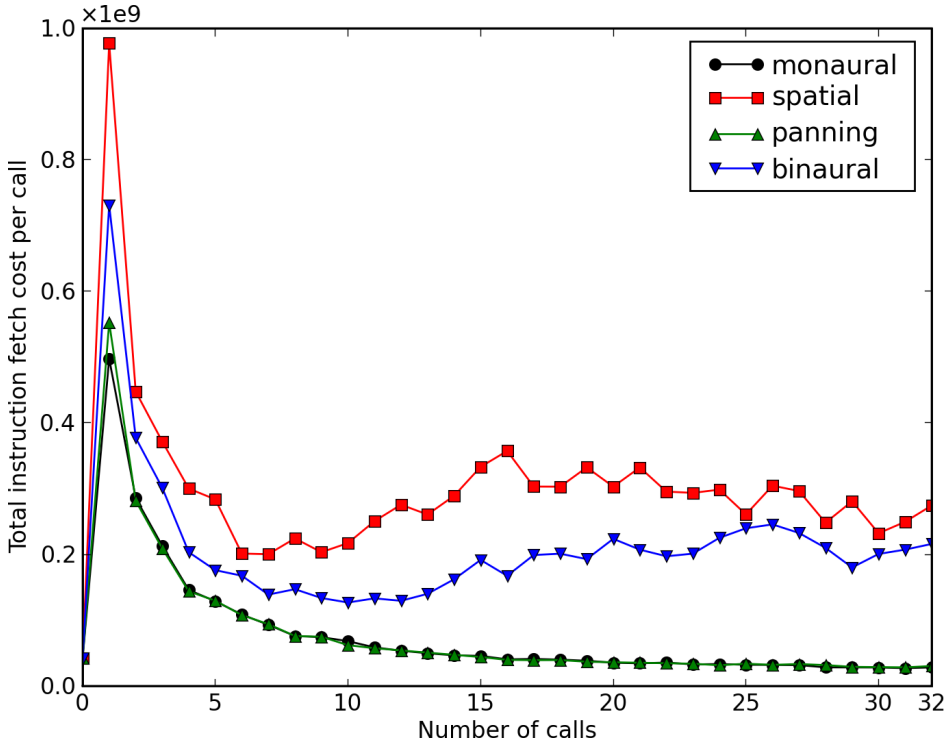


Figure 6.23: *Total instruction cost per call for each presentation mode.*

6.12 Conclusion

The results from the experiments performed in the preceding chapter and the use of AMT to carry out these experiments will now be discussed.

6.12.1 Spatial Audio Experiments

Two experiments were conducted to determine the effect of spatial audio on speaker identification in multiple source situations the first with four possible sources and the second with six. The first experiment, given in Section 6.7 on page 81 only compared monaural audio to HRTF spatialisation. The experiment found that, on average, subjects are able to correctly identify the active source 43% of the time with monaural audio and 88% of the time with spatialisation using HRTFs. The experiment demonstrated that the system gives an approximately two-fold increase in the probability of the user correctly identifying the active speaker in a multiple speaker situation where one of four different speakers could be active at any given time. The experiment was conducted using subjects that were not familiar with the system or the specific HRTFs used in the experiment. Correct identification rates are expected to improve after the user becoming more familiar with the HRTFs [82] or when using a set of HRTFs from a subject with anthropometric measurements similar to their own. The second speaker identification experiment is described in Section 6.8 on page 83. Although the absolute scores for this experiment are lower than those of the first experiment, a similar trend exists, with the three spatial audio models affording a listener an approximately two-fold increase in correctly identifying the active speaker. The decrease in absolute identification rates compared to the earlier experiment is most likely due to the greater number of possible source and the use of three spatial audio models being more likely to confuse the listener. Headphone panning has more dramatic energy distribution than the other spatial audio, with a source panned completely to the left presenting all the energy to the left year, which can explain why headphone panning outperforms the HRTF spatialisation and the basic binaural model in speaker identification tasks.

The speech intelligibility experiment performed in Section 6.9 on page 88 demonstrated that spatial audio outperforms monaural audio in all cases when there is more than one competing speaker. Additionally, the HRTF spatialisation in some cases provided the listener greater intelligibility with one more masking speaker than monaural audio. The HRTF spatialisation performed the best in the four masker situation, a difficult scenario due to the high level of energetic masking because of the noise masker. This shows that while the panning and binaural models provide better than monaural results, the use of a more realistic acoustic model that is measured from the physical world and not constructed based on approximations performs the best in extreme situations. In some cases, HRTF spatialisation rendered more intelligible audio than the monaural model did in a situation with one fewer source.

The audio encoding effect experiment performed in Section 6.10 on page 93 demonstrated

that spatial audio delivers better results when the bandwidth of the audio is increased, but that this improvement is only comes close to being proportional to the bit rate increase for HRTF spatialisation. Spatialisation handles a decrease in bit rate better than monaural audio in some cases and worse than monaural audio in others, a general trend has not yet been observed. More research is required to better understand the mechanism by which compression effects spatialisation.

The benchmarking done in Section 6.11 on page 100 gives an indication of the increased computational complexity faced when implementing real-time audio spatialisation. The headphone panning model, which is by far the least complex of the three, requires only 3% more Central Processing Unit (CPU) power than monaural audio. HRTF spatialisation and the basic binaural model are the most expensive, using 803% and 586% more CPU power than monaural audio respectively. The basic binaural model requires 75.97% of the power necessary for HRTF spatialisation.

The choice of which spatial audio to use will depend on the circumstances of the problem that needs to be solved. The basic binaural model fares the worst of the three, it does not perform as well as any of the other spatial models but requires so much as 76% of the CPU power of the HRTF model. The performance of the binaural mode can likely be increased if a different processing method than the FIR filter approach borrowed from the HRTF spatialisation mode is adopted. However, regardless of computational performance, the binaural does not function as well as the computationally inexpensive panning mode. The experiments performed in this chapter demonstrate spatial audio is an improvement over monaural audio in multiple speaker situations. The experiments also show that the choice of spatial model will depend on the application and the resources available for the application. HRTF spatialisation provides the best intelligibility in situations with competing speakers, albeit at a great computational cost. If only speaker identification is required, then headphone panning is the clear winner, in both performance and low resource usage. In situations where intelligibility is important, HRTF spatialisation makes the best choice unless available resources force the use of the panning model.

6.12.2 AMT Experimentation

It is possible to use AMT to conduct human-subject experiments as long as a great deal of care is put into careful experiment design. The greatest advantage of using AMT is that a large number of subjects can be gotten to do an experiment in a relatively short time. The data obtained from AMT is unfortunately not very reliable and bad data will need to be scrubbed before usage. Controlled test runs need to be dispersed within the actual test runs so that individuals who are attempting to chance the system and are not completing the tests correctly can be identified. A significant portion of the collected data needed to be discarded. The low validity rate of the data will also increase the financial cost of any experiments.

6.13 Summary

Psychoacoustic experiments were conducted both on students at Stellenbosch University as well as on the AMT web service. The anonymity that AMT affords workers leads to many of the subjects not performing the experiment correctly. If data collected from experiments conducted using AMT are to be used in research, then many controls and checks need to be built into the experiment to ensure that the invalid data can be safely and easily discarded.

The experiments were conducted over the Internet so that a larger audience could be reached than would be possible in the available time frame if subjects had to be brought in to do the experiment under supervision. An unsupervised experiment does run the risk of cheaters polluting the collected data. Controls need to be dispersed within the experiment so that subjects with suspicious results can be removed from the data set. A generic website framework for performing auditory experiments over the Internet was created. The website was built in HTML and PHP with the audio samples played using Adobe Flash. The subjects' answers were saved in a MySQL database for later processing.

Two experiments were performed to ascertain the effect of spatial audio on speaker identification, the first and earlier experiment was administered by hand and the second using the experiment website. The first had the subjects attempt to identify the active speaker out of four possibles and compared HRTF spatialisation to monaural audio. The second featured six possible speakers that could be active and added headphone panning and binaural audio to the list of auditory presentation modes. The next experiment that was performed set out to measure the intelligibility of a target speaker in in the presence of concurrent masking speakers, for each of the four presentation modes. The number of masking speakers ranged from one to four, with the fourth speaker being a white noise signal to make for an especially difficult final scenario. The aim of the final psychoacoustic experiment was to determine the effect of lossy audio encoding and compression on spatialisation, that is to see if intelligibility suffers more or less when encoding the audio than it does with monaural audio. The final experiment set out to measure the computational cost of each acoustic model relative to that of monaural audio.

The experiments demonstrated that spatial audio affords greater speech intelligibility and assists users in identifying the active speaker in multiple speaker situations, at the costs of additional CPU usage.

Chapter 7

Conclusion

The primary hypothesis of our research is that implementing a spatial audio model (in place of the monaural audio model that is traditionally deployed in electronic voice-based communication) will improve the user's communication experience.

This chapter concludes the work in the preceding chapters of this thesis. Suggestions for future work are made. Remarks about the success of the project as a whole are made at the end of this chapter.

7.1 Future Work

The research, although self contained, is an initial foray into a large research field where much work can still be done. The remainder of this section will give details on future work within the scope of this project.

7.1.1 Improvement of Spatialisation Accuracy

A key component to the performance of the system in terms of presenting the correct perceived spatial location of the audio source to the user is dependent on the quality of the HRTFs that are used. Generalised HRTFs from a downloadable database were used for the system and are non-individualised because they were not measured using the subject's own auditory system. Research shows that such non-individualised HRTFs can lead to errors in localisation, especially with respect to elevation and front-to-back discrimination [137, 135]. Performance can be improved by using HRTFs that are either measured directly from the subject [36], which is impractical due to the complex measurement procedure [36, 72], or by individualising generalised HRTFs using anthropometric measurements [138, 84]. A functional model can make the HRTFs independent of subject [139].

7.1.2 Increasing Performance

For the system to be deployed to hardware systems with lower specifications, the performance of the system with respect to CPU usage will need to be increased. If the high order

FIR filters used for HRTF spatialisation were instead approximated by lower order Infinite Impulse Response (IIR) filters performance would be increased greatly [78, 85, 77]. Performance could also be improved if the complexity of the FIR filters can be decreased without sacrificing the fidelity of the spatial images.

High performance is also of utmost importance for successful usage of spatial audio in a virtual world. Such a situation will typically have a client need to spatialise a large number of audio streams due to the large number of avatars that might exist in the region, which will be taxing on the client. A source clustering algorithm, wherein far away and unimportant sources in close proximity are mixed together and spatialised as one stream, will go a ways towards achieving this.

7.1.3 Further Experimentation

Significantly more extensive experimentation than was possible in Chapter 6 on page 71, due to time constraints, still needs to be done. Spatial audio, being a perceptual phenomenon, necessitates that human-subject experiments be performed for the validation of any hypotheses regarding improvements to the user's experience. Some possible experiments that could still be conducted,

- determining the effect of training and familiarity with spatial audio on localisation,
- measuring the relative merits and demerits of different HRTF databases,
- changes in virtual world immersion as a result of using spatial audio and
- determining if the addition of reverberation simulation is beneficial or not.

This section also ties in with the future work mentioned in Section 7.1.2, experimentation with reduced complexity HRTF models will give insight into how far the model can be simplified while still retaining accurate spatialisation.

More experiments need to be done regarding the effect of audio encoding and compression, with different voice codecs, to determine the cases in which spatial audio will suffer the least degradation. The speech codecs that function best for monaural audio might not necessarily be the best for spatial audio and the best codec for spatial audio needs to be found.

7.2 Final Remarks

The monaural audio model that is used for electronic communication has not changed since the telephone was first conceived and a paradigm shift to a model that uses spatial audio will greatly change how we perceive and use audio in electronic communication. We believe that this project contributes to the successful design and implementation of such a model.

The experiments that were performed support the research hypothesis that spatial audio is beneficial to users in situations with more than one possible speaker, or situations with

more than one concurrent speaker. Spatial audio was demonstrated to afford a listener an approximately two-fold increase in speaker identification for both four and six speaker scenarios over monaural audio. The headphone panning model performed the best out of the three spatial audio models. Spatial audio resulted in greater speech intelligibility of a target speaker than monaural audio in all scenarios with more than once competing speaker. The HRTF spatialisation model even delivered more intelligible audio in some situations than the monaural model did in a situation with one fewer masker.

Out of the three spatial audio models, the headphone panning and HRTF spatialisation models are the clear winners. The panning model performs the best in source identification tasks and is by far the least computationally expensive of the three. The HRTF model provides the best intelligibility gains, but at a much greater computational cost than headphone panning. The binaural model does not perform as well as the other two and is not significantly cheaper than HRTF spatialisation so as to make it worthwhile to implement. As a general recommendation, HRTF spatial is best when high fidelity spatial cues are necessary and when resources are abundant and headphone panning is best in situations with severely limited resources. The client-side approach that was adopted requires all processing to be done by the client. The processing load placed on the client can be reduced by instead using a server-side approach, which is costly to implement, or a hybrid approach, which sacrifices flexibility.

In closing, this research has proven that spatial audio can be implemented in a modern telephony system and that such an implementation is an improvement over the monaural audio model currently in use, when engaging in conversations with multiple participants.

Bibliography

[1] http://interface.cipic.ucdavis.edu/CIL_tutorial/3D_HRTF/HRTF_hor.htm. Last accessed: November, 2009.

[2] "Amazon Mechanical Turk." https://www.mturk.com/mturk/welcome. Last accessed: November, 2009.

[3] "Diamondware data sheet." http://www.dw.com/dev/archives/flyers/DW_Datasheet.pdf. Last accessed: November, 2009.

[4] "GNU profiler." http://www.cs.utah.edu/dept/old/texinfo/as/gprof_toc.html. Last accessed: November, 2009.

[5] "LibOpenMetaverse." http://www.openmetaverse.org/projects/libopenmetaverse. Last accessed: November, 2009.

[6] "Linden Scripting Language." http://wiki.secondlife.com/wiki/LSL_Portal. Last accessed: November, 2009.

[7] "Listen HRTF database." http://recherche.ircam.fr/equipes/salles/listen/. Last accessed: November, 2009.

[8] "OpenSim." http://opensimulator.org/wiki/Main_Page. Last accessed: November, 2009.

[9] "OpenSim REST support." http://opensimulator.org/wiki/REST. Last accessed: November, 2009.

[10] "OpenSim Web Statistics Module." http://opensimulator.org/wiki/Web_Statistics_Module. Last accessed: November, 2009.

[11] "Performance testing." http://wiki.aqsis.org/dev/performance_testing. Last accessed: November, 2009.

[12] "PJMEDIA reference: Media ports framework." `http://www.pjsip.org/pjmedia/docs/html/group__PJMEDIA__PORT.htm`. Last accessed: November, 2009.

[13] "PJSIP." `http://www.pjsip.org/docs.htm`. Last accessed: November, 2009.

[14] "PJSIP media flow." `http://trac.pjsip.org/repos/wiki/media-flow`. Last accessed: November, 2009.

[15] "Pjsua manual." `http://www.pjsip.org/pjsua.htm`. Last accessed: November, 2009.

[16] "Second Life." `http://secondlife.com`. Last accessed: November, 2009.

[17] "Second life llappviewer class documentation." `http://doc.daleglass.net/ll/release/d6/df6/llappviewer_8cpp.html`. Last accessed: November, 2009.

[18] "Second Life quaternions." `http://wiki.secondlife.com/wiki/Quaternion`. Last accessed: November, 2009.

[19] "Skype business." `http://www.skype.com/business`. Last accessed: November, 2009.

[20] "Speex website." `http://www.speex.org`. Last accessed: November, 2009.

[21] "SRS CS headphone." `http://www.srslabs.com/content.aspx?id=427`. Last accessed: November, 2009.

[22] "Suzuki Laboratory HRTF database." `http://www.ais.riec.tohoku.ac.jp/lab/db-hrtf/index.html`. Last accessed: November, 2009.

[23] "Through my ears." `http://throughmyears.blogspot.com`. Last accessed: November, 2009.

[24] "Valgrind." `http://valgrind.org`. Last accessed: November, 2009.

[25] "Virtual barbershop." `http://ccgi.bluerabbit.plus.com/virtualbarbershop`. Last accessed: November, 2009.

[26] "Perfect surround cinema sound without loudspeakers." `http://www.beyerdynamic.de/index.php?id=1466&&L=1&tx_ttnews[tt_news]=280&tx_ttnews[backPid]=121&cHash=df0612e41a`, July 2007. Last accessed: November, 2009.

[27] "PJSIP: Doing it in stereo." `http://blog.pjsip.org/2008/03/31/doing-it-in-stereo`, March 2008. Last accessed: November, 2009.

[28] "Internet growth in south africa." `http://www.internetworldstats.com/af/za.htm`, December 2009. Last accessed: December, 2009.

[29] "OECD broadband portal." `http://www.oecd.org/sti/ict/broadband`, May 2009. Last accessed: November, 2009.

[30] "Vodacom broadband." `http://www.vodacom.co.za/services/vodacom_broadband/index.jsp`, 2009. Last accessed: November, 2009.

[31] ADELSON, E. H. and BERGEN, J. R., "The plenoptic function and the elements of early vision." in *Computational Models of Visual Processing*, pp. 3–20, MIT Press, 1991.

[32] ADOBE, "Flash Player penetration." `http://www.adobe.com/products/player_census/flashplayer`, September 2009. Last accessed: November, 2009.

[33] AJDLER, T., FALLER, C., SBAIZ, L., and VETTERLI, M., "Sound Field Analysis Along a Circle and its Applications to HRTF Interpolation." *Journal of the Audio Engineering Society*, March 2008, Vol. 56, No. 3, pp. 156–175.

[34] AJDLER, T., SBAIZ, L., and VETTERLI, M., "The Plenacoustic Function and Its Sampling." *IEEE Transactions on Signal Processing*, October 2006, Vol. 54, No. 10, pp. 3790–3804.

[35] AKEROYD, M. A., "The psychoacoustics of binaural hearing." *International Journal of Audiology*, 2006, Vol. 45, No. Supplement 1, No. Supplement 1, pp. 25–33.

[36] ALGAZI, V., DUDA, R., THOMPSON, D., and AVENDANO, C., "The CIPIC HRTF Database." in *Proc. 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp. 99–102, 2001.

[37] ALLEN, J. B. and BERKLEY, D. A., "Image method for efficiently simulating small-room acoustics." *The Journal of the Acoustical Society of America*, April 1979, Vol. 65, No. 4, pp. 943–950.

[38] ARAKI, S., SAWADA, H., and MAKINO, S., "Blind speech separation in a meeting situation with maximum SNR beamformers." in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007*, vol. 1, pp. 41–45, 2007.

[39] ARBOGAST, T. L., MASON, C. R., and KIDD, J. G., "The effect of spatial separation on informational and energetic masking of speech." *The Journal of the Acoustical Society of America*, November 2002, Vol. 112, No. 5, pp. 2086–2098.

[40] ARNOLD, J., "Voip on the verge." `http://www.telecoms-mag.com/Americas/article.asp?HH_ID=AR_539`, November 2004. Last accessed: November, 2009.

[41] ARONS, B., "A Review of the Cocktail Party Effect." *Journal of the American Voice I/O Society*, 1992, Vol. 12, pp. 35–50.

[42] ARRINGTON, M., "Modeling the real market value of social networks." `http://www.techcrunch.com/2008/06/23/modeling-the-real-market-value-of-social-networks`, June 2008. Last accessed: November, 2009.

[43] ARRINGTON, M., "New lawsuit brings clarity to skype's ip problem (prognosis: Screwed)." `http://www.techcrunch.com/2009/09/18/new-lawsuit-brings-clarity-to-skypes-ip-problem`, September 2009. Last accessed: November, 2009.

[44] ATWOOD, J., "3d positional audio and hrtfs." `http://www.codinghorror.com/blog/archives/000494.html`, January 2006. Last accessed: November, 2009.

[45] AU, W. J., "Second life takes aim at skype." `http://gigaom.com/2009/05/19/second-life-takes-aim-at-skype`, May 2009. Last accessed: November, 2009.

[46] BALDIS, J. J., "Effects of spatial audio on memory, comprehension, and preference during desktop conferences." in *CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems*, (New York, NY, USA), pp. 166–173, ACM, 2001.

[47] BEGAULT, D., *3-D Sound for Virtual Reality and Multimedia*. Academic Press, 1994.

[48] BELL, A. G., "Improvement in telegraphy." US patent 174,465, March 1876.

[49] BERKHOUT, A. J., DE VRIES, D., and VOGEL, P., "Acoustic control by wave field synthesis." *The Journal of the Acoustical Society of America*, May 1993, Vol. 93, No. 5, pp. 2764–2778.

[50] BERKHOUT, A., "A holographic approach to acoustic control." *Journal of the Audio Engineering Society*, December 1988, Vol. 36, No. 12, pp. 977–995.

[51] BEST, V., CARLILE, S., JIN, C., and VAN SCHAIK, A., "The Role of High Frequencies in Speech Localization." *The Journal of the Acoustical Society of America*, 2005, Vol. 118, pp. 353–363.

[52] BIRCHFIELD, S. and GANGISHETTY, R., "Acoustic localization by interaural level difference." in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05)*, vol. 4, 2005.

[53] BLAUERT, J., *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, 1997.

[54] BRONKHORST, A. W., "The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multiple-Talker Conditions." *Acta Acustica united with Acustica*, January 2000, Vol. 86, No. 1, pp. 117–128.

[55] BRUNGART, D. S. and RABINOWITZ, W. M., "Auditory localization of nearby sources. Head-related transfer functions." *The Journal of the Acoustical Society of America*, September 1999, Vol. 106, No. 3, pp. 1465–1479.

[56] BYRNE, D., DILLON, H., TRAN, K., ARLINGER, S., WILBRAHAM, K., COX, R., HAGERMAN, B., HETU, R., KEI, J., LUI, C., KIESSLING, J., KOTBY, M. N., NASSER, N. H. A., KHOLY, W. A. H. E., NAKANISHI, Y., OYER, H., POWELL, R., STEPHENS, D., MEREDITH, R., SIRIMANNA, T., TAVARTKILADZE, G., FROLENKOV, G. I., WESTERMAN, S., and LUDVIGSEN, C., "An international comparison of long-term average speech spectra." *The Journal of the Acoustical Society of America*, October 1994, Vol. 96, No. 4, pp. 2108–2120.

[57] CASTRO, E., "Bye bye embed." `http://www.alistapart.com/articles/byebyeembed`, July 2006. Last accessed: November, 2009.

[58] CHERRY, E. C., "Some Experiments on the Recognition of Speech, with One and with Two Ears." *The Journal of the Acoustical Society of America*, 1953, Vol. 25, No. 5, No. 5, pp. 975–979.

[59] CHRISTENSEN, F., JENSEN, C., and MOLLER, H., "The design of VALDEMAR – an artificial head for binaural recording purposes." in *Proceedings of the Audio Engineering Society's 109th Convention*, (Los Angeles, California, USA), September 2000. (preprint 5253).

[60] COOKE, M., BARKER, J., CUNNINGHAM, S., and SHAO, X., "An audio-visual corpus for speech perception and automatic speech recognition." *The Journal of the Acoustical Society of America*, November 2006, Vol. 120, No. 5, pp. 2421–2424.

[61] CRINON, R. J., KHAN, H. M., and KUKOLECA, D., "Active speaker identification." US patent application US20080312923, December 2008.

[62] DEMPSEY, M. J., "Method for introducing harmonics into an audio stream for improving three-dimensional audio positioning." US patent 6,215,879, April 2001.

[63] DINAN, M., "Report: International skype calls increased 41 percent in 2008." `http://ip-pbx.tmcnet.com/topics/ip-pbx/articles/52855-report-international-skype-calls-increased-41-percent-2008.htm`, March 2009. Last accessed: November, 2009.

[64] DIPAOLA, S. and COLLINS, D., "A 3D Virtual Environment for Social Telepresence." in *Proceedings of the Western Computer Graphics Symposium*, Citeseer, March 2002.

[65] DOLCI, W., "Abgradcon 2009: A glimpse into mixed-reality meetings of the future." `http://astrobiology.nasa.gov/articles/abgradcon-2009-a-glimpse-into-mixed-reality-meetings-of-the-future`, July 2009. Last accessed: November, 2009.

[66] DRULLMAN, R. and BRONKHORST, A. W., "Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation." *The Journal of the Acoustical Society of America*, April 2000, Vol. 107, No. 4, pp. 2224–2235.

[67] EVEREST, F. A., *Master Handbook of Acoustics*. 4 edition. McGraw-Hill Professional, 2000.

[68] Federal Communications Commission, "Comment sought on transition from circuit-switched network to all-ip network." `http://hraunfoss.fcc.gov/edocs_public/attachmatch/DA-09-2517A1.pdf`, December 2009. Last accessed: November, 2009.

[69] FIELDING, R. T., *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine, 2000.

[70] FRENCH, N. R. and STEINBERG, J. C., "Factors Governing the Intelligibility of Speech Sounds." *The Journal of the Acoustical Society of America*, January 1947, Vol. 19, No. 1, pp. 90–119.

[71] FREYMAN, R. L., HELFER, K. S., MCCALL, D. D., and CLIFTON, R. K., "The role of perceived spatial separation in the unmasking of speech." *The Journal of the Acoustical Society of America*, December 1999, Vol. 106, No. 6, pp. 3578–3588.

[72] GARDNER, W. G. and MARTIN, K. D., "HRTF measurements of a KEMAR." *The Journal of the Acoustical Society of America*, June 1995, Vol. 97, No. 6, pp. 3907–3908.

[73] GOLDMAN, A., "IBM touts 3D environment as superior to IM." `http://blog.internetnews.com/agoldman/2009/06/ibm-sametime-3d-im.html`, June 2009. Last accessed: November, 2009.

[74] GREENWOOD, D. D., "A cochlear frequency-position function for several species—29 years later." *The Journal of the Acoustical Society of America*, June 1990, Vol. 87, No. 6, pp. 2592–2605.

[75] GRIESINGER, D., "Stereo and Surround Panning in Practice." in *Audio Engineering Society 112th Convention*, pp. 1–6, 2002.

[76] HAMMERSHOI, D. and MOLLER, H., "Methods for binaural recording and reproduction." *Acta Acustica united with Acustica*, 2002, Vol. 88, No. 3, No. 3, pp. 303–311.

[77] HASEGAWA, H., KASUGA, M., MATSUMOTO, S., and KOIKE, A., "Binaural sound reproduction using head-related transfer functions (HRTFs) approximated by IIR filters." in *TENCON 99. Proceedings of the IEEE Region 10 Conference*, vol. 1, pp. 150–153, 1999.

[78] HASEGAWA, H., KASUGA, M., MATSUMOTO, S., and KOIKE, A., "Simply realization of sound localization using HRTF approximated by IIR filter." *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, June 2000, Vol. 83, No. 6, pp. 973–978.

[79] HAWLEY, M. L., LITOVSKY, R. Y., and COLBURN, H. S., "Speech intelligibility and localization in a multi-source environment." *The Journal of the Acoustical Society of America*, June 1999, Vol. 105, No. 6, pp. 3436–3448.

[80] HEIDE, D. A. and KANG, G. S., "Speech Enhancement for Bandlimited Speech." in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 1, pp. 393–396 vol.1, May 1998.

[81] HO, K. and SUN, M., "Passive source localization using time differences of arrival and gain ratios of arrival." *IEEE Transactions on Signal Processing*, February 2008, Vol. 56, No. 2, pp. 464–477.

[82] HOFMAN, P. M., RISWICK, J. G. V., and OPSTAL, A. J. V., "Relearning sound localization with new ears." *Nature Neuroscience*, September 1998, Vol. 1, No. 5, pp. 417–421.

[83] HOLMAN, T., *5.1 Surround Sound: Up and Running*. Focal Press, 2000.

[84] HU, H., ZHOU, L., MA, H., and WU, Z., "HRTF personalization based on artificial neural network in individual virtual auditory space." *Applied Acoustics*, July 2008, Vol. 69, No. 2, pp. 163 – 172.

[85] HUOPANIEMI, J. and KARJALAINEN, M., "Comparison of Digital Filter Design Methods for 3-D Sound." in *in Proc. IEEE Nordic Signal Processing Symp. (NORSIG'96*, pp. 131–134, 1996.

[86] IBM, "Made in IBM labs: IBM creates software for holding face-to-face meetings in virtual worlds." `http://www-03.ibm.com/press/us/en/pressrelease/26837.wss`, May 2009. Last accessed: November, 2009.

[87] JAKOBSSON, M., "Experimenting on mechanical turk." Palo Alto Research Centre blog, `http://blogs.parc.com/blog/2009/07/experimenting-on-mechanical-turk-5-how-tos/`, July 2009. Last accessed: November, 2009.

[88] JEAN-MARC VALIN, "The Speex codec manual." `http://www.speex.org/docs/manual/speex-manual.pdf`, December 2007. Last accessed: November, 2009.

[89] JIN, C., SCHENKEL, M., and CARLILE, S., "Neural system identification model of human sound localization." *The Journal of the Acoustical Society of America*, September 2000, Vol. 108, No. 3, pp. 1215–1235.

[90] JONES, S. and FOX, S., "Generations online in 2009." Pew Internet and American Life Project, `http://pewinternet.org/Reports/2009/Generations-Online-in-2009.aspx`, 2009. Last accessed: November, 2009.

[91] KAHRS, M. and BRANDENBURG, K., *Applications of digital signal processing to audio and acoustics*. Kluwer Academic Publishers, 1998.

[92] KANADA, Y., "Multi-Context Voice Communication Controlled By using An Auditory Virtual Space." in *Conference on Communication and Computer Networks*, 2004.

[93] KANADA, Y., "Multi-context voice communication in a SIP/SIMPLE-based shared virtual sound room with early reflections." in *NOSSDAV '05: Proceedings of the international workshop on Network and operating systems support for digital audio and video*, (New York, NY, USA), pp. 45–50, ACM, 2005.

[94] KEEGAN, V., "Virtual worlds are getting a second life." `http://www.guardian.co.uk/technology/2009/jul/29/virtual-worlds`, July 2009. Last accessed: November, 2009.

[95] KIDD, G., JR, ARBOGAST, T. L., MASON, C. R., and GALLUN, F. J., "The advantage of knowing where to listen." *The Journal of the Acoustical Society of America*, December 2005, Vol. 118, No. 6, pp. 3804–3815.

[96] KING, R. B. and OLDFIELD, S. R., "The Impact of Signal Bandwidth on Auditory Localization: Implications for the Design of Three-Dimensional Audio Displays." *Human Factors*, June 1997, Vol. 39, No. 2, pp. 287–296.

[97] KITTUR, A., CHI, E. H., and SUH, B., "Crowdsourcing User Studies With Mechanical Turk." in *Proceedings of the 26th Annual ACM Conference on Human Factors in Computing Systems (CHI2008); 2008 April 5-10; Florence, Italy*, pp. 453–465, 2008.

[98] KOMINEK, J. and BLACK, A. W., "The CMU Arctic speech databases." in *Fifth ISCA Workshop on Speech Synthesis*, ISCA, June 2004.

[99] KORZENIOWSKI, P., "Three technologies you need in 2009." Forbes, `http://www.forbes.com/2009/01/08/small-business-voip-ent-tech-cx_bm_0108bmightytech09.html`, January 2009. Last accessed: November, 2009.

[100] KUIPERS, J. B., *Quaternions and Rotation Sequences: A Primer with Applications to Orbits, Aerospace and Virtual Reality*. Princeton University Press, 2002.

[101] LAINE, M., "WordPress Audio Player." `http://wpaudioplayer.com`. Last accessed: November, 2009.

[102] LAMONT, I., "Second Life claims social network crown." `http://www.pcworld.com/article/172691/second_life_claims_social_network_crown.html`, September 2009. Last accessed: November, 2009.

[103] Linden Lab, "1 billion hours, 1 billion dollars served: Second Life celebrates major milestones for virtual worlds." `http://lindenlab.com/pressroom/releases/22_09_09`, September 2009. Last accessed: November, 2009.

[104] MACMANUS, R., "Report: Enterprise virtual worlds more effective than web conferencing." `http://www.readwriteweb.com/archives/enterprise_virtual_worlds.php`, December 2008. Last accessed: November, 2009.

[105] MAES, C., "Why profile?." Lectures Notes, `www.stanford.edu/class/cme212/profiling.pdf`, February 2008. Last accessed: November, 2009.

[106] MATSUMOTO, M., YAMANAKA, S., TOYAMA, M., and NOMURA, H., "Effect of Arrival Time Correction on the Accuracy of Binaural Impulse Response Interpolation: Interpolation Methods of Binaural Response." *Journal of the Audio Engineering Society*, February 2004, Vol. 52, No. 1/2, pp. 56–61.

[107] MINDY MCADAMS, "Embedded MP3 audio player." `http://www.macloo.com/examples/audio_player`. Last accessed: November, 2009.

[108] MINNAAR, P., OLESEN, S., CHRISTENSEN, F., and MOLLER, H., "Localization with binaural recordings from artificial and human heads." *Journal of the Audio Engineering Society*, May 2001, Vol. 49, No. 5, pp. 323–336.

[109] MURRAY, J., ERWIN, H., and WERMTER, S., "Robotics Sound-Source localization and Tracking Using Interaural Time Difference and Cross-Correlation." in *AI Workshop on NeuroBotics*, Citeseer, 2004.

[110] NAEF, M., STAADT, O., and GROSS, M., "Spatialized audio rendering for immersive virtual environments." in *VRST '02: Proceedings of the ACM symposium on Virtual reality software and technology*, (New York, NY, USA), pp. 65–72, ACM, 2002.

[111] NISHINO, T., MASE, S., KAJITA, S., TAKEDA, K., and ITAKURA, F., "Interpolating HRTF for auditory virtual reality." *The Journal of the Acoustical Society of America*, October 1996, Vol. 100, No. 4, pp. 2602–2602.

[112] NTT DOCOMO, "Docomo develops spatial audio transmission technology for mobile phones." `http://www.nttdocomo.com/pr/2009/001438.html`, May 2009. Last accessed: November, 2009.

[113] OELLERS, H., "Wave field synthesis and holophony." `http://www.syntheticwave.de`, 2009. Last accessed: November, 2009.

[114] PEISSIG, J. and KOLLMEIER, B., "Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners." *The Journal of the Acoustical Society of America*, March 1997, Vol. 101, No. 3, pp. 1660–1670.

[115] PROAKIS, J. G. and MANOLAKIS, D. G., *Digital Signal Processing: Principles, Algorithms and Applications*. 4 edition. Prentice Hall, 2006.

[116] RÄMÖ, A. and TOUKOMAA, H., "On comparing speech quality of various narrow- and wideband speech codecs." in *Proceedings of the Eighth International Symposium on Signal Processing and Its Applications, 2005*, vol. 2, pp. 603–606, August 2005.

[117] ROMAN, N. and WANG, D., "Binaural Tracking of Multiple Moving Source." *IEEE Transactions on Audio, Speech, and Language Processing*, May 2008, Vol. 16, No. 4, pp. 728–739.

[118] ROSENBERG, J., SCHULZRINNE, H., CAMARILLO, G., JOHNSTON, A., PETERSON, J., SPARKS, R., HANDLEY, M., and SCHOOLER, E., "SIP: Session initiation protocol." RFC 3261, 2002. Last accessed: November, 2009.

[119] RUMSEY, F., *Spatial Audio*. Focal Press, 2001.

[120] SARGENT, E. W., HERRMANN, B., HOLLENBEAK, C. S., and BANKAITIS, A. E., "The minimum speech test battery in profound unilateral hearing loss." *Otology & Neurotology*, July 2001, Vol. 22, No. 4, pp. 480–486.

[121] SHEPARD, R. N., "Circularity in Judgments of Relative Pitch." *The Journal of the Acoustical Society of America*, December 1964, Vol. 36, No. 12, pp. 2346–2353.

[122] SHINN-CUNNINGHAM, B. G., "Spatial hearing advantages in everyday environments." in *Proceedings of Office of Naval Research workshop on Attention, Perception, and Modeling for Complex Displays*, (Troy, New York), June 2003.

[123] SINGH, K. and SCHULZRINNE, H., "Peer-to-peer internet telephony using SIP." in *NOSSDAV '05: Proceedings of the international workshop on Network and operating systems support for digital audio and video*, (New York, NY, USA), pp. 63–68, ACM, 2005.

[124] SINNREICH, H. and JOHNSTON, A. B., *Internet Communications Using SIP: Delivering VoIP and Multimedia Services with Session Initiation Protocol*. 2 edition. Wiley, 2006.

[125] SODNIK, J., SUSNIK, R., and TOMAZIC, S., "Resolution enhancement of a general HRTF library." in *Proceedings of ACOUSTICS*, 2005.

[126] SPORS, S., RABENSTEIN, R., and AHRENS, J., "The Theory of Wave Field Synthesis Revisited." in *124th AES Convention*, (Amsterdam, The Netherlands), May 2008.

[127] Strategy Analytics, Inc, "Virtual worlds forecast to grow at 23% through 2015." `http://www.strategyanalytics.com/default.aspx?mod=PressReleaseViewer&a0=4745`, 2009. Last accessed: November, 2009.

[128] TELTSCHER, S., GRAY, V., MAGPANTAY, E., VALLEJO, I., MANIEWICZ, M., and WOODALL, M., "Information Society Statistical Profiles 2009 – Afrcia." tech. rep., International Telecommunication Union (ITU), 2009.

[129] VELTMAN, J. A., OVING, A. B., and BRONKHORST, A. W., "3-D Audio in the Fighter Cockpit Improves Task Performance." *International Journal of Aviation Psychology*, 2004, Vol. 14, pp. 239 – 256.

[130] VERMAAK, J. and BLAKE, A., "Nonlinear filtering for speaker tracking in noisy and reverberantenvironments." in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01)*, vol. 5, 2001.

[131] WALLING, S., "How to: Use virtual worlds for business." `http://www.readwriteweb.com/enterprise/2009/07/how-to-use-virtual-worlds-for-business.php`, July 2009. Last accessed: November, 2009.

[132] WARD, D. and WILLIAMSON, R., "Particle filter beamforming for acoustic source localization in areverberant environment." in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002. Proceedings.(ICASSP'02)*, vol. 2, 2002.

[133] WAUTERS, R., "Skype opens up to sip, finally eyes enterprise customers the way it should." `http://www.techcrunch.com/2009/03/23/skype-opens-up-to-sip-finally-eyes-enterprise-customers-the-way-it-should`, March 2009. Last accessed: November, 2009.

[134] WENTK, R., "The evolution of virtual worlds." `http://www.techradar.com/news/gaming/the-evolution-of-virtual-worlds-610257`, June 2009. Last accessed: November, 2009.

[135] WENZEL, E. M., ARRUDA, M., KISTLER, D. K., and WIGHTMAN, F. L., "Localization using nonindividualized head-related transfer functions." *Journal of the Acoustic Society of Amera*, July 1993, Vol. 94(1), pp. 111–123.

[136] WILLIAMS, E. G., *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*. Academic Press, 1999.

[137] XU, S., LI, Z., and SALVENDY, G., "Individualization of Head-Related Transfer Function for Three-Dimensional Virtual Auditory Display: A Review." in *Virtual Reality*, pp. 397–407, Springer, Berlin / Heidelberg, 2007.

[138] XU, S., LI, Z., and SALVENDY, G., "Improved method to individualize head-related transfer function using anthropometric measurements." *Acoustical Science and Technology*, November 2008, Vol. 29, No. 6, pp. 388–390.

[139] ZHANG, W., KENNEDY, R., and ABHAYAPALA, T., "Efficient Continuous HRTF Model Using Data Independent Basis Functions: Experimentally Guided Approach." *Audio, Speech, and Language Processing, IEEE Transactions on*, May 2009, Vol. 17, No. 4, pp. 819–829.

Appendix A

Nomenclature

A.1 Acronyms

AMR Adaptive Multi-Rate

AMR-WB Adaptive Multi-Rate Wideband

AMT Amazon Mechanical Turk

CPU Central Processing Unit

FIR Finite Impulse Response

HIT Human Intelligence Task

HRIR Head-Related Impulse Response

HRTF Head-Related Transfer Function

HTML Hyper Text Markup Language

HTTP Hypertext Transfer Protocol

IAX Inter-Asterisk Exchange

IIR Infinite Impulse Response

ILD Interaural Level Difference

IP Internet Protocol

IRC Internet Relay Chat

ITD Interaural Time Difference

KEMAR Knowles Electronic Manikin for Acoustic Research

LAN Local Area Network

LSL	Linden Scripting Language
MMOG	Massively Multiplayer Online Game
MOS	Mean Opinion Score
MSE	Mean Squared Error
MySQL	My Structured Query Language
PCM	Pulse-Code Modulation
PDF	Probability Density Function
PHP	PHP: Hypertext Preprocessor
PSTN	Public Switched Telephone Network
REST	Representational State Transfer
RTCP	RTP Control Protocol
RTP	Real-time Transport Protocol
SIP	Session Initiation Protocol
SNR	Signal-to-Noise Ratio
TOA	Time of Arrival
UDP	User Datagram Protocol
URI	Uniform Resource Identifier
UUID	Universally Unique Identifier
VoIP	Voice over Internet Protocol
XML-RPC	Extensible Markup Language – Remote Procedure Call

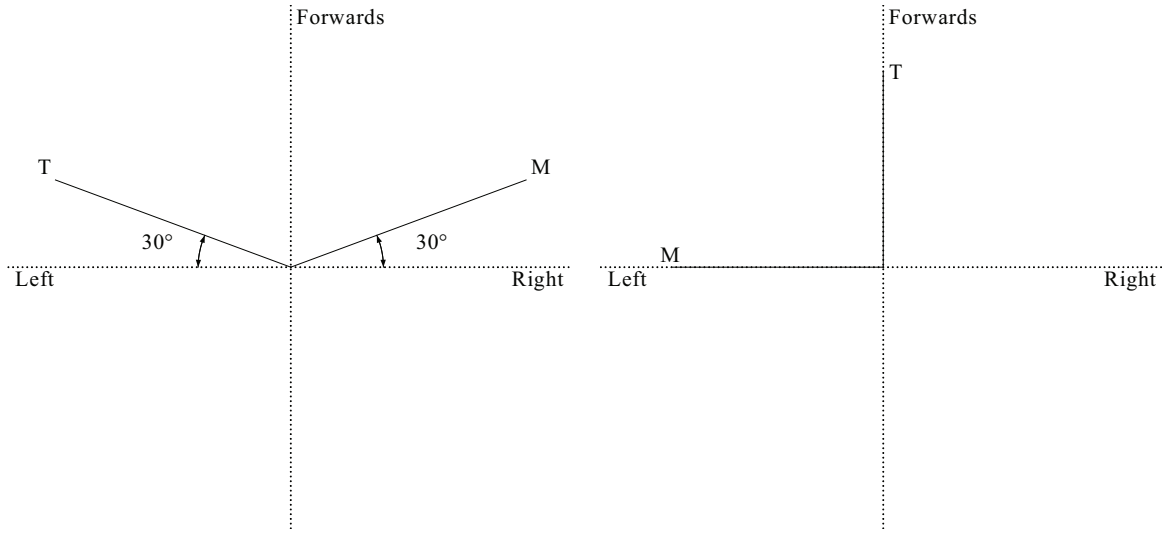
Appendix B

Experimental Setup

This chapter provides additional details on the setup of the experiments conducted in Chapter 6 on page 71 that is not necessary for interpretation of the results but would assist in further experiments.

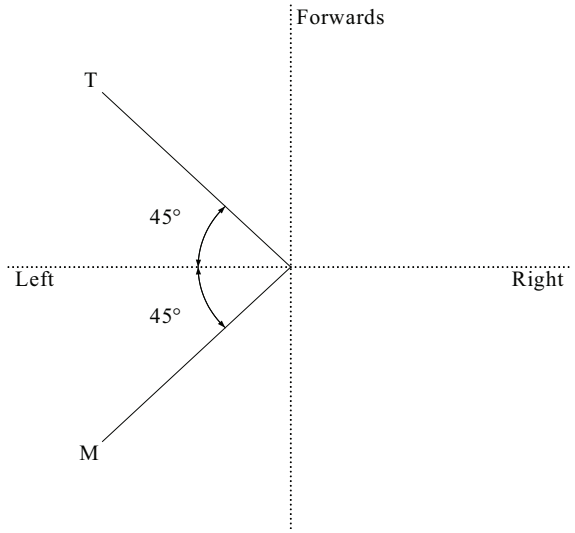
B.1 Positional Scenarios

This section gives the target and maskers positions for the experiment conducted in Section 6.9 on page 88.
Table B.1shows the azimuth positions of the target and masker sources for each of the sixteen different scenarios, where N is the number of maskers. Figures B.1, B.2, B.3 and B.4 show the target and maskers position, where T designates a target, M a speech masker and W a noise masker.

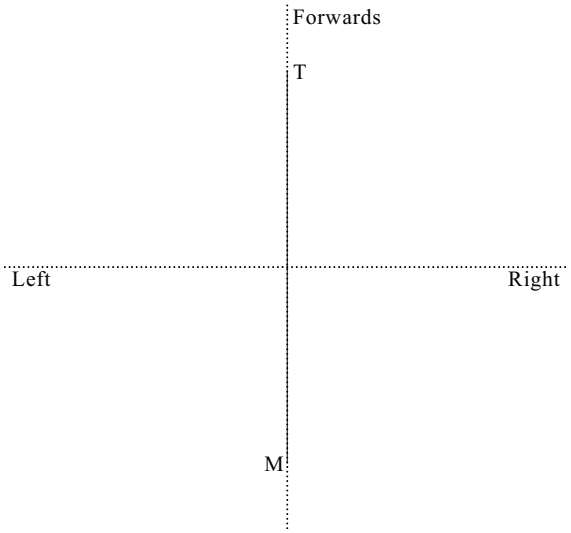


(a) Scenario 1

(b) Scenario 2

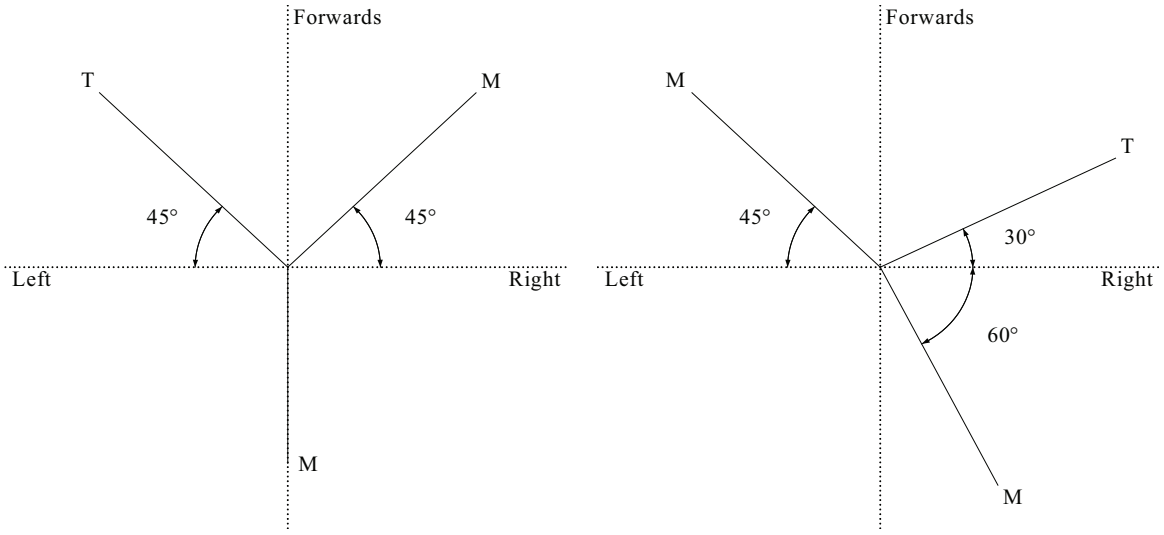


(c) Scenario 3



(d) Scenario 4

Figure B.1: *Source positions for scenarios with one masker, where T designates a target and M a speech masker.*



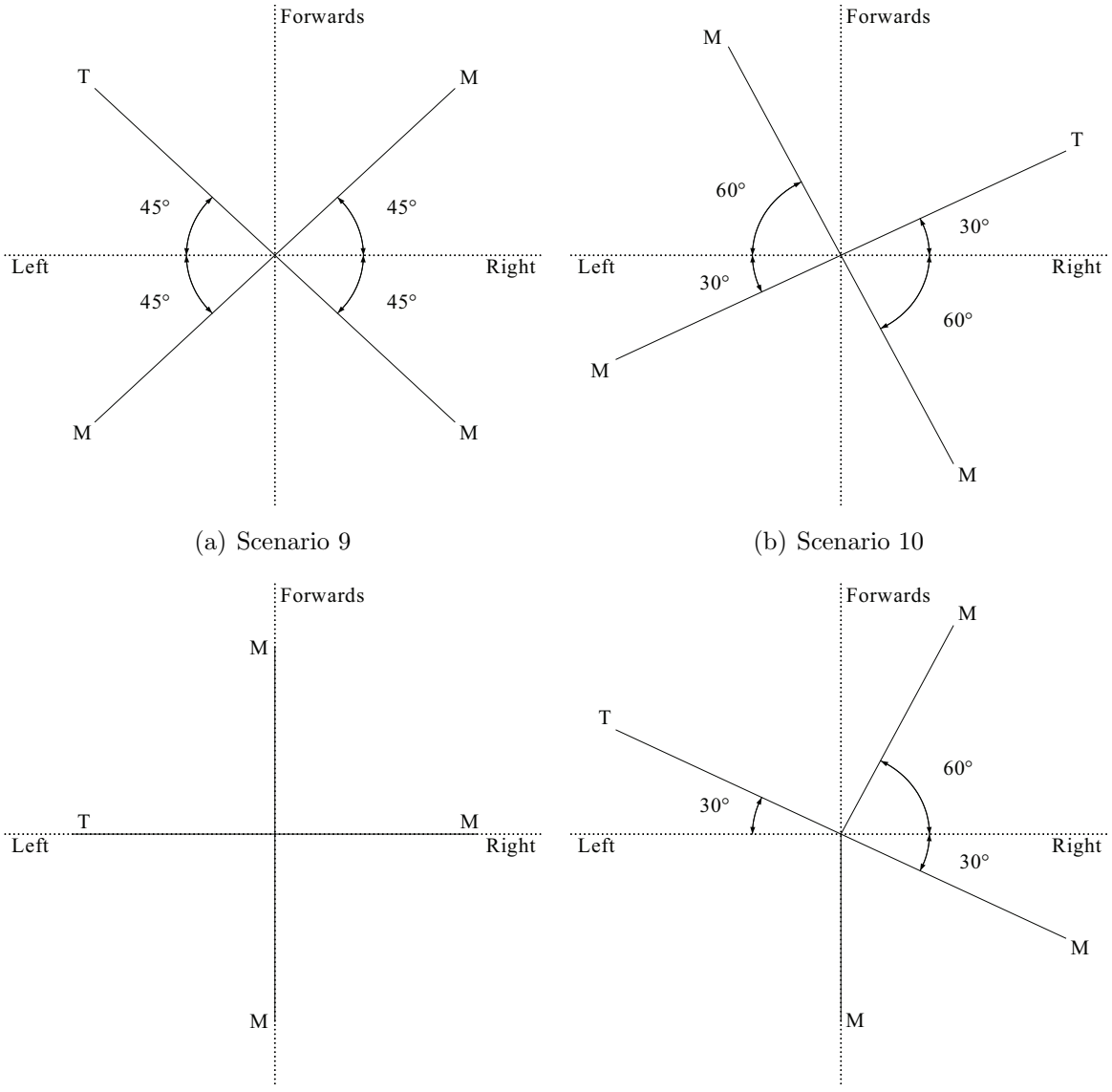
(a) Scenario 5

(b) Scenario 6

(c) Scenario 7

(d) Scenario 8

Figure B.2: *Source positions for scenarios with two maskers, where T designates a target and M a speech masker.*



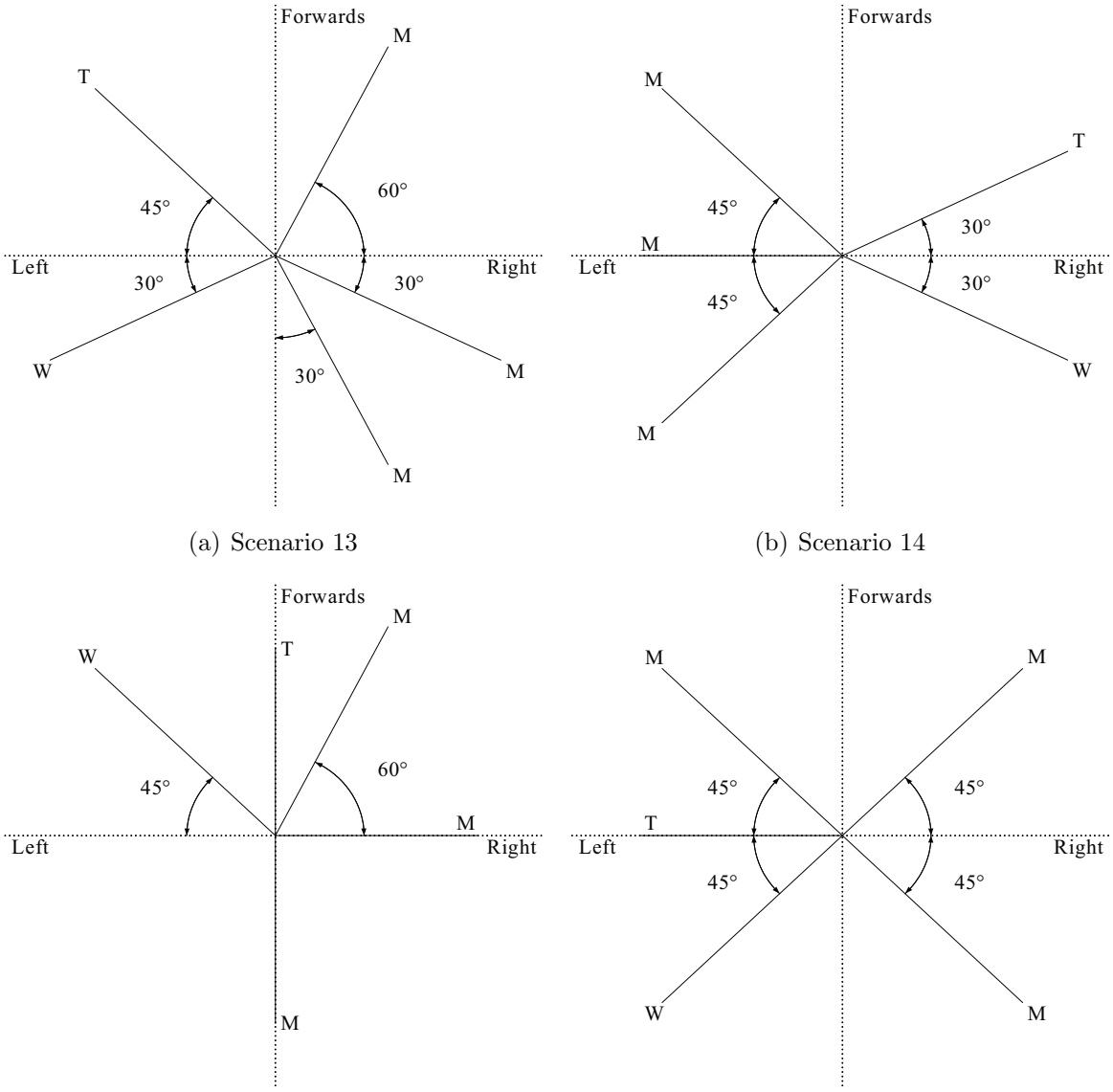
(a) Scenario 9

(b) Scenario 10

(c) Scenario 11

(d) Scenario 12

Figure B.3: *Source positions for scenarios with three maskers, where T designates a target and M a speech masker.*



(a) Scenario 13

(b) Scenario 14

(c) Scenario 15

(d) Scenario 16

Figure B.4: *Source positions for scenarios with four maskers, where T designates a target, M a speech masker and W a noise masker.*

Table B.1: *Azimuth positions of the target and masker sources. N is the number of maskers.*

Scenario	N	Target	Masker 1	Masker 2	Masker 3	Masker 3
1	1	330°	60°	-	-	-
2	1	0°	270°	-	-	-
3	1	315°	225°	-	-	-
4	1	0°	180°	-	-	-
5	2	315°	45°	180°	-	-
6	2	60°	150°	315°	-	-
7	2	315°	120°	225°	-	-
8	2	0°	120°	240°	-	-
9	3	315°	45°	135°	225°	-
10	3	60°	150°	240°	330°	-
11	3	270°	0°	90°	180°	-
12	3	300°	30°	120°	180°	-
13	4	315°	30°	120°	150°	240°
14	4	60°	225°	270°	315°	120°
15	4	0°	30°	90°	180°	315°
16	4	270°	345°	135°	315°	225°